Dædalus

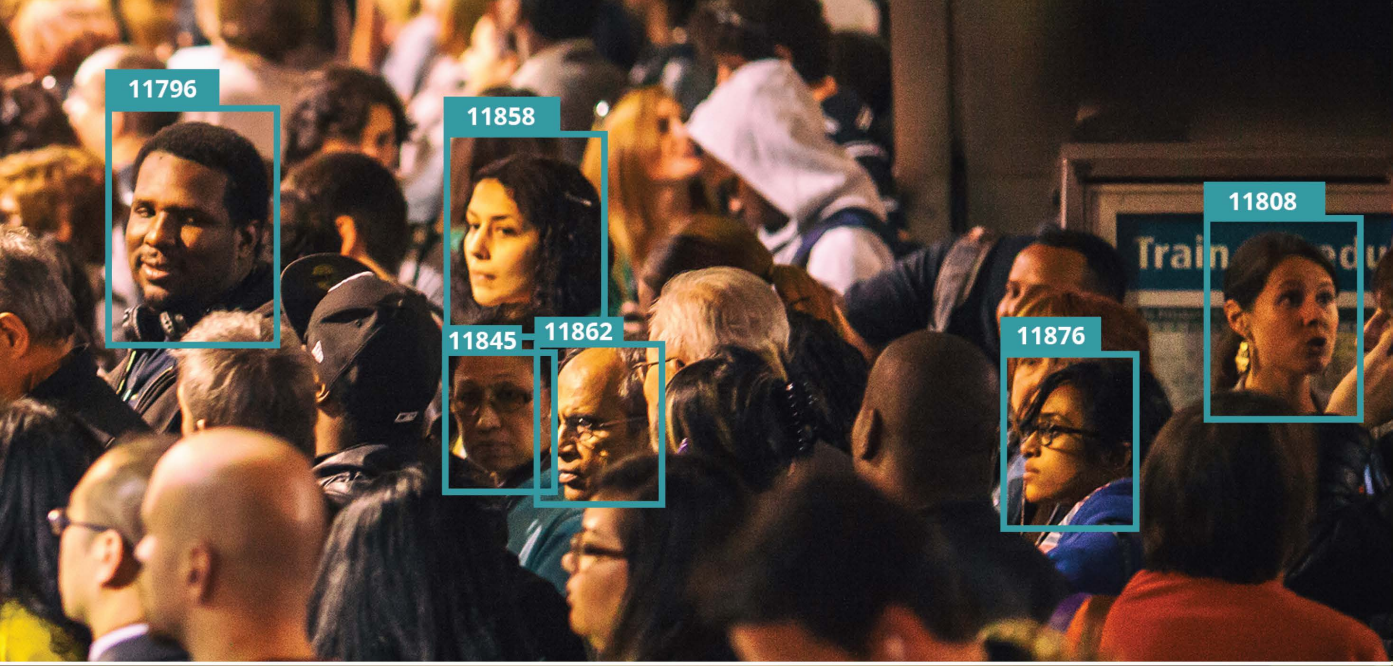
Journal of the American Academy of Arts & Sciences

Winter 2024

Understanding Implicit Bias: Insights & Innovations

Goodwin Liu & Camara Phyllis Jones, guest editors

with David Baltimore · David S. Tatel
Anne-Marie Mazza · Eric H. Holder, Jr.
Marcella Nunez-Smith · Kirsten N. Morehouse
Mahzarin R. Banaji · Kate A. Ratliff
Colin Tucker Smith · Andrew N. Meltzoff
Walter S. Gilliam · Jennifer T. Kubota
Manuel J. Galvan · B. Keith Payne · Rebecca C. Hetey
MarYam G. Hamedani · Hazel Rose Markus
Jennifer L. Eberhardt · Anthony G. Greenwald
Thomas Newkirk · Jack Glaser · Jerry Kang
Alexandra Kalev · Frank Dobbin · Wanda A. Sigur
Nicholas M. Donofrio · Alice Xiang · Darren Walker
Thomas D. Albright · William A. Darity Jr.
Diana Dunn · Rayid Ghani · Deena Hayes-Greene
Tanya Katerí Hernández · Sheryl Heron


11796
11858
11845
11862
11876
11808


HIGH SCHOOL
1319
EAST 41ST STREET
LOS ANGELES SCHOOL POLICE
SERGEANT

Dædalus

Journal of the American Academy of Arts & Sciences

"Understanding Implicit Bias: Insights & Innovations"
Volume 153, Number 1; Winter 2024

Goodwin Liu & Camara Phyllis Jones, Guest Editors
Phyllis S. Bendell, Editor in Chief
Peter Walton, Associate Editor
Key Bird, Assistant Editor
Maya Robinson, Assistant Editor

Inside front cover: (top) **Artificial intelligence.** Facial recognition technology identifies human faces on a BART (Bay Area Rapid Transit) platform during evening rush hour in San Francisco, California. Photograph © 2012 by Thomas Hawk. Image published under a Creative Commons Attribution-NonCommercial 2.0 Generic (CC BY-NC 2.0 DEED) license. Image modified.

(middle) **Health care.** Brenda Major (left) is examined by Dr. Fernanda Mercade during a routine checkup at the Jessie Trice Center for Community Health clinic in Miami, Florida. Photograph © 2012 by Joe Raedle/Getty Images.

(bottom) **Schools.** Los Angeles School Police Sergeant Robert Carlborn watches students line up to pass through a security checkpoint at Thomas Jefferson High School in Los Angeles, California. Photograph © 2005 by David McNew/Getty Images.

Contents

Dædalus

Journal of the American Academy of Arts & Sciences

The labyrinth designed by Dædalus for King Minos of Crete, on a silver tetradrachma from Cnossos, Crete, c. 350–300 BC (35 mm, Cabinet des Médailles, Bibliothèque Nationale, Paris). "Such was the work, so intricate the place, / That scarce the workman all its turns cou'd trace; / And Dædalus was puzzled how to find / The secret ways of what himself design'd."
–Ovid, *Metamorphoses, Book 8*

Dædalus was founded in 1955 and established as a quarterly in 1958. Its namesake was renowned in ancient Greece as an inventor, scientist, and unriddler of riddles. The journal's emblem, a labyrinth seen from above, symbolizes the aspiration of its founders to "lift each of us above his cell in the labyrinth of learning in order that he may see the entire structure as if from above, where each separate part loses its comfortable separateness."

The American Academy of Arts & Sciences, like its journal, brings together distinguished individuals from every field of human endeavor. It was chartered in 1780 as a forum "to cultivate every art and science which may tend to advance the interest, honour, dignity, and happiness of a free, independent, and virtuous people." Now in its third century, the Academy, with its more than five thousand members, continues to provide intellectual leadership to meet the critical challenges facing our world.

Dædalus Winter 2024
Issued as Volume 153, Number 1

Editorial offices: *Dædalus*, American Academy of Arts & Sciences, 136 Irving Street, Cambridge MA 02138. Phone: 617 576 5085. Fax: 617 576 5088. Email: daedalus@amacad.org.

Library of Congress Catalog No. 12-30299.

Dædalus (ISSN 0011-5266; E-ISSN 1548-6192) is published quarterly (winter, spring, summer, fall) by The MIT Press, 255 Main Street Suite 9, Cambridge MA 02142, for the American Academy of Arts & Sciences. An electronic full-text version of *Dædalus* is available from amacad.org/daedalus and from The MIT Press.

Members of the American Academy please direct all questions and claims to daedalus@amacad.org.

Advertising inquiries may be addressed to Marketing Department, MIT Press Journals, 255 Main Street Suite 9, Cambridge MA 02142. Phone: 617 253 2866. Fax: 617 253 1709. Email: journals-info@mit.edu.

Printed in the United States by The Sheridan Press, 450 Fame Avenue, Hanover PA 17331.

Postmaster: Send address changes to *Dædalus*, 255 Main Street Suite 9, Cambridge MA 02142. Periodicals postage paid at Boston MA and at additional mailing offices.

The typeface is Cycles, designed by Sumner Stone at the Stone Type Foundry of Guinda CA. Each size of Cycles has been separately designed in the tradition of metal types.

Preface: Recognizing Implicit Bias in the Scientific & Legal Communities

David Baltimore, David S. Tatel & Anne-Marie Mazza

Several years ago, in the Fall 2018 volume of *Dædalus*, we wrote "Bridging the Science-Law Divide," an essay about the work of the National Academies of Sciences, Engineering, and Medicine's Committee on Science, Technology, and Law.[1] In that essay, we discussed the importance of having the legal and scientific communities engage with each other on a host of issues, and highlighted work that the committee conducted on the courts' handling of scientific evidence and on society's governance of emerging technologies. We mentioned that, in the coming years, the committee hoped to focus on the issue of implicit bias (referred to as "unconscious bias" in our 2018 essay), as it was becoming increasingly evident that factors outside individual awareness were affecting personal and institutional decision-making that hindered the full participation of all our citizens.

In a provocative talk at Georgetown University in 2017, Justice Ruth Bader Ginsburg remarked that confronting unconscious bias would be the next big challenge for the courts. The Supreme Court, in an opinion by Justice Anthony M. Kennedy, recognized the importance of addressing disparate impact liability, as it helps uncover discriminatory intent and counteract unconscious prejudices.[2]

Not only are the courts wrestling with implicit bias but society has begun to recognize that implicit bias is a challenge for society at large, playing out in all kinds of environments: education, policing, housing, and everyday activities. In facing this challenge, we have been thrilled to receive encouragement from colleagues like Darren Walker, president of the Ford Foundation, who agreed to support our effort to focus on the science of implicit bias by providing our committee with the opportunity to organize a workshop on this important topic. The 2021 workshop, entitled "The Science of Implicit Bias: Implications for Law and Policy," which was thoughtfully cochaired by Justice Goodwin Liu and Dr. Camara Jones, vividly highlighted how implicit bias is hindering our country's ability to give all citizens opportunities to reach their full potential, and become fully engaged members of our nation.

As we see from the essays in this volume – that focus on what science tells us about implicit bias, what the implications of not addressing it are for a fair and equitable society, and what might be done to lessen its impact – implicit bias does not have to be the determining factor in our decision-making. We can build a so-

https://doi.org/10.1162/daed_e_02043

ciety and institutions that take steps to mitigate some of its harmful effects. Thus, we hope you find the essays in this collection informative. We were delighted to read pieces by many of the experts who participated in the 2021 workshop and to learn from others who agreed to contribute to this volume.

ABOUT THE AUTHORS

David Baltimore, a Fellow of the American Academy since 1974, is the Judge Shirley Hufstedler Professor of Biology and President Emeritus at the California Institute of Technology. He is the author of over seven hundred articles in virology and immunology.

David S. Tatel, a Fellow of the American Academy since 2015, is Senior Counsel at Hogan Lovells, and a retired Judge of the United States Court of Appeals for the District of Columbia Circuit.

Anne-Marie Mazza is the Senior Director of the Committee on Science, Technology, and Law for the National Academies of Sciences, Engineering, and Medicine, and former Executive Director of the President's Council of Advisors on Science and Technology.

ENDNOTES

[1] David Baltimore, David S. Tatel, and Anne-Marie Mazza, "Bridging the Science-Law Divide," *Dædalus* 147 (4) (Fall 2018): 181–194, https://www.amacad.org/publication/bridging-science-law-divide.

[2] *Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc.*, 576 U.S. 519 (2015).

Introduction: Implicit Bias in the Context of Structural Racism

Goodwin Liu & Camara Phyllis Jones

In September of 1967, with the civil rights movement in full stride, Dr. Martin Luther King Jr. gave a major address at the annual meeting of the American Psychological Association. In that speech, Dr. King sought to enlist the help of "members of the academic community, who are constantly writing about and dealing with the problems that we face and who have the tremendous responsibility of molding the minds of young men and women all over the country."[1] He called for deeper understanding of the nation's legacy of racism and said "the understanding needs to be carefully documented and consequently more difficult to reject." He urged social scientists to "'tell it like it is'" – to illuminate why "the Negro, after 350 years of handicaps, mired in an intricate network of contemporary barriers, [can]not be ushered into equality by tentative and superficial changes." Racial oppression, he said, arose from "systemic" causes and "will [not] be solved until there is a kind of cosmic discontent enlarging in the bosoms of people of good will all over this nation."

Since Dr. King's time, social scientists and other scholars have contributed enormously to our understanding of inequality based on race, gender, and other lines of socially constructed difference. One finding of this body of work is that although overt expressions of racism and other prejudices have declined over several decades, unequal opportunities and outcomes persist in education, employment, housing, health care, the justice system, and other domains. The causes are complex and varied and cannot be reduced to a single explanation. But one thing we know is that racial and other biases have not been extirpated and continue to reinforce these inequalities. Even as overt prejudice has decreased, implicit bias – the associations we make automatically, outside of our conscious awareness, between certain groups and certain characteristics – is a prominent feature of ordinary cognition that can impair our ability to treat people fairly despite our best intentions. The strength and pervasiveness of implicit bias pose a major challenge to actualizing our societal commitment to equality.

This issue of *Dædalus* features state-of-the-art research and insightful perspectives on implicit bias from a variety of disciplines and domains. The authors in-

https://doi.org/10.1162/daed_a_02044

clude many of the leading scholars on this topic, as well as prominent policymakers with deep experience navigating issues of diversity and discrimination. The volume serves as an up-to-date compendium of the literature and identifies directions for further study. It is an invaluable resource for anyone interested in the current state of knowledge about implicit bias, its causes and effects, and potential interventions and mitigation strategies.

The genesis of this volume was a workshop we led on the science of implicit bias convened by the Committee on Science, Technology, and Law of the National Academies of Sciences, Engineering, and Medicine in March 2021. The event was also guided by an interdisciplinary planning group, some of whom have offered their perspectives in this volume. The workshop, held online during the early stages of the coronavirus pandemic, drew more than one thousand people worldwide and featured some of the cutting-edge research in these pages. When we started planning the workshop in early 2020, we could not have foreseen so many recent events relevant to this work.

The murder of George Floyd by Minneapolis police officer Derek Chauvin in May 2020, caught on video, has ignited a national and global movement to combat anti-Black racism. The pandemic, together with racialized scapegoating, has fueled a sharp rise in anti-Asian hate incidents and violence, including the killing of six Asian American women in Atlanta in March 2021 just days before our workshop. We have also witnessed barely disguised racism in anti-immigration rhetoric by public officials and commentators. Even as the Supreme Court endorsed colorblindness in its 2023 decision ending affirmative action in university admissions, it blinks reality to ignore that race continues to shape key aspects of people's lives today. Racism denial underlies much contemporary rhetoric and motivates many policy decisions across our nation. Moreover, despite progress in education and other areas, women's equality remains elusive in many domains of public and private life, with unique challenges for women of different races, to say nothing of prejudice and open hostility directed at transgender people. Legislatures, courts, corporations, universities, K–12 schools, and organizations at all levels are earnestly grappling with these issues, and as was true in Dr. King's time, there is an urgent need for scholarship that can illuminate these challenges and possible solutions.

Each essay in this volume conveys important findings and ideas that merit careful consideration on their own. Collectively, the essays highlight three themes we find especially significant. First, thanks to three decades of research, the existence of implicit bias as a demonstrable and observable reality rests on a firm and wide-ranging evidence base. Since 1998, over thirty million Implicit Association Tests have been taken, measuring unconscious or implicit attitudes and stereotypes on a variety of dimensions, including race, gender, age, reli-

gion, sexual orientation, weight, and others. The results comprise a large dataset that shows the extent of implicit biases in favor of advantaged groups as well as changes over time at a societal level.[2] In addition, careful studies from a variety of disciplines, including psychology, sociology, economics, law, and neuroscience, have reported powerful evidence of implicit bias through laboratory experiments, audit studies, other field studies, brain imaging techniques, and, most recently, research on natural language processing.[3]

These studies have demonstrated the operation of implicit bias not only in laboratory tasks but also in real-life decision-making in education, employment, health care, the justice system, commercial transactions, and even sports. There is disturbing evidence of such bias in law enforcement and voting.[4] And some of the most poignant work has revealed how young children acquire racial biases from their observations of adult interactions, suggesting that such biases can be "caught" at an early age, even when not explicitly taught, and transmitted across generations.[5]

As many of the authors note, research shows that the correlation between implicit bias and discriminatory behavior is small to moderate, and we must be careful to examine all the facts before ascribing any individual incident, such as an employment decision or a police shooting, to implicit bias. But even small correlations between predispositions and behaviors add up over an individual's lifetime and at the level of society-wide decisions and interactions.[6] Consider, in this regard, the growing evidence that geographic regions with higher levels of implicit bias tend to have higher levels of racial disparities across a number of socially significant outcomes, such as law enforcement, education, and health care.[7] These findings and others have bolstered an emerging view that implicit bias may be understood as a feature of groups or geographic places, not just individual minds.[8]

The overarching point is that thirty years of scientific inquiry has produced a compelling body of evidence demonstrating the existence, strength, and pervasiveness of implicit bias. The societal challenges posed by this body of research are serious and cannot be ignored.

Second, while much of the foundational research on implicit bias has come from psychology, a prominent theme of emerging work focuses on the relationship between implicit bias and structural inequality. The plethora of studies revealing how our biases manifest outside of conscious awareness have made fascinating contributions to the science of cognition. But these studies should not be understood to "psychologize" racism or other biases – as if such biases exist solely or primarily as the mental states of individuals – just as neuroscientific study of implicit bias should not be construed as "biologizing racism."[9] Implicit bias resides within a larger context of systemic discrimination, whereby laws, policies, and institutional practices assign value or allocate opportunity in ways that advantage certain groups and disadvantage others across multiple domains.[10] Implicit bias is both a cause and an effect of structural inequalities.

How else to explain the remarkable finding that the extent of slaveholding by county at the time of Abraham Lincoln's presidency correlates with county-level measures of pro-white implicit bias today, even after controlling for self-reported attitudes?[11] A natural inference is that this relationship is mediated by structural inequalities – including de jure and de facto segregation, wealth and education gaps, disparate treatment by the justice system, and more – that have maintained racial hierarchy across generations. "Chronic exposure to these structural inequalities maintains and exacerbates implicit bias."[12] Moreover, as noted, recent research has found that regional differences in implicit racial bias are correlated with the extent of racial disparities in policing, educational, health, and economic outcomes. It seems all but certain that the arrow of causation runs in both directions.

This point is also brought home by emerging studies of bias in artificial intelligence (AI). Because AI reflects the patterns that exist in its training data, it is not surprising that a variety of algorithms – from facial recognition to health care utilization to public safety risk evaluation – exhibit racial bias in their output and decision-making.[13] In addition, recent work on massive language corpora (that is, the entirety of language in certain formats or repositories on the internet, such as Google Books) has demonstrated how implicit racial and gender biases in individual minds can amount to a reservoir of "collectively held or culturally imprinted beliefs" that complicate the task of ensuring algorithmic fairness in training AI.[14] In all these ways, our understanding is becoming more clear that implicit bias is not simply a matter of individual beliefs and attitudes, but also an expression and enabler of structural inequality at an institutional and societal level.

Third, compared to the robust research demonstrating the existence and operation of implicit bias, the evidence base for effective interventions or mitigation strategies is still emerging. We expect that it will continue to develop further in the coming years. A key question is whether implicit bias is malleable and can be lessened in individuals through various forms of priming, education, or other contextual interventions. The available evidence provides scant reason to believe that durable change can be achieved through modest interventions, including some current forms of diversity or implicit bias training.[15] This is unsurprising given the extent to which our implicit biases reflect mental associations reinforced through a lifetime of observations and stimuli, starting from an early age.[16] At the same time, there is evidence that implicit racial bias at a societal level has decreased over the past fifteen years, and more research is needed to understand what conditions facilitated such change.[17]

For a number of reasons, antibias training of the kind often used by corporations, universities, and other organizations not only shows little promise for changing bias or behavior over the long term, but also has the potential to backfire.[18] Instead of efforts to "debias" individual minds, changing organizational

policies and structures appears to be necessary to prevent or counteract the operation of implicit bias and to create new patterns of interaction that reflect our expressed commitments to fairness, inclusion, and equal opportunity. Such changes may require strong leadership with clearly stated values, along with strategies to promote intergroup contact under conditions in which people of different backgrounds work together as equals toward a common goal.[19] Combatting implicit bias may also require changes in antidiscrimination law and judicial interpretations, as well as structural or procedural reforms that reduce discretion in decision-making.[20]

The emerging picture is one in which implicit bias, though grounded in cognitive science, is increasingly being understood as a phenomenon that both maintains and manifests systemic inequalities with long histories and structural underpinnings. As aptly stated in this volume with regard to race :

> Conceptualizing implicit racial bias as merely a byproduct of human cognition overlooks the critical scientific insight that racial bias exists not only in the head, but also in the world. Implicit bias is the residue that an unequal world leaves on an individual's mind and brain, residue that has been created and built into institutional policies and practices and socialized into patterns of behavior over hundreds of years through the workings of culture.[21]

Accordingly, it is unlikely that implicit bias can be effectively addressed by cognitive interventions alone, without broader institutional, legal, and structural reforms. Such reforms may require organizations to collect data, analyze disparities, and take concrete and sustained actions to root out inequitable practices.[22] They will require individuals and organizations throughout society to acknowledge that, despite their stated values or best intentions, their current ways of doing things – including existing diversity, equity, and inclusion initiatives – are not immune to implicit bias and may not be sufficient to prevent its operation or remedy its effects. All of this is hard work, but it is necessary and urgent work if we are to counter implicit bias in its individual and systemic dimensions.

We are indebted to the many scholars and leaders who have contributed to this volume. Their knowledge provides critical insights into how far we still have to go to achieve a just and equitable society, and how we might take steps to get there. We are also grateful to Anne-Marie Mazza, Steven Kendall, the National Academies' Committee on Science, Technology, and Law, as well as Phyllis Bendell, her talented staff, and the American Academy of Arts and Sciences for their dedication to this important topic and for facilitating the work of our authors. We are honored to bring you this issue of *Dædalus*.

ABOUT THE AUTHORS

Goodwin Liu, a Fellow of the American Academy since 2019 and Chair of the Board of Directors of the American Academy since 2022, is an Associate Justice of the California Supreme Court. He is the author of *Keeping Faith with the Constitution* (with Pamela S. Karlan and Christopher H. Schroeder, 2010) and numerous articles on inequality in public education and diversity in higher education and the legal profession.

Camara Phyllis Jones, a Fellow of the American Academy since 2022, is a Leverhulme Visiting Professor in Global Health and Social Medicine at King's College London, an Adjunct Professor at the Rollins School of Public Health at Emory University, and a Senior Fellow and Adjunct Associate Professor at the Morehouse School of Medicine. She is the editor of *Black Women and Resilience: Power, Perseverance, and Public Health* (with Kisha Braithwaite Holden, 2024) and many articles on naming, measuring, and addressing the impacts of racism on the health and well-being of the United States and the world.

ENDNOTES

[1] "King's Challenge to the Nation's Social Scientists," *The APA Monitor* 30 (1) (1999), https://www.apa.org/topics/equity-diversity-inclusion/martin-luther-king-jr-challenge.

[2] Kirsten N. Morehouse and Mahzarin R. Banaji, "The Science of Implicit Race Bias: Evidence from the Implicit Association Test," *Dædalus* 153 (1) (Winter 2024): 21–50, https://www.amacad.org/publication/science-implicit-race-bias-evidence-implicit-association-test; and Kate A. Ratliff and Colin Tucker Smith, "The Implicit Association Test," *Dædalus* 153 (1) (Winter 2024): 51–64, https://www.amacad.org/publication/implicit-association-test.

[3] Rebecca C. Hetey, MarYam G. Hamedani, Hazel Rose Markus, and Jennifer L. Eberhardt, "'When the Cruiser Lights Come On': Using the Science of Bias & Culture to Combat Racial Disparities in Policing," *Dædalus* 153 (1) (Winter 2024): 123–150, https://www.amacad.org/publication/when-cruiser-lights-come-using-science-bias-culture-combat-racial-disparities-policing; Jennifer T. Kubota, "Uncovering Implicit Racial Bias in the Brain: The Past, Present & Future," *Dædalus* 153 (1) (Winter 2024): 84–105, https://www.amacad.org/publication/uncovering-implicit-racial-bias-brain-past-present-future; and Morehouse and Banaji, "The Science of Implicit Race Bias."

[4] Eric H. Holder, Jr., "Seeing the Unseen," *Dædalus* 153 (1) (Winter 2024): 15–17, https://www.amacad.org/publication/seeing-unseen.

[5] Andrew N. Meltzoff and Walter S. Gilliam, "Young Children & Implicit Racial Biases," *Dædalus* 153 (1) (Winter 2024): 65–83, https://www.amacad.org/publication/young-children-implicit-racial-biases.

[6] Jerry Kang, "Little Things Matter a Lot: The Significance of Implicit Bias, Practically & Legally," *Dædalus* 153 (1) (Winter 2024): 193–212, https://www.amacad.org/publication/little-things-matter-lot-significance-implicit-bias-practically-legally.

[7] Morehouse and Banaji, "The Science of Implicit Race Bias."

[8] Ratliff and Smith, "The Implicit Association Test"; and Manuel J. Galvan and B. Keith Payne, "Implicit Bias as a Cognitive Manifestation of Systemic Racism," *Dædalus* 153 (1)

(Winter 2024): 106–122, https://www.amacad.org/publication/implicit-bias-cognitive-manifestation-systemic-racism.

[9] Kubota, "Uncovering Implicit Racial Bias in the Brain," 95.

[10] Galvan and Payne, "Implicit Bias as a Cognitive Manifestation of Systemic Racism"; and Hetey, Hamedani, Markus, and Eberhardt, "'When the Cruiser Lights Come On.'"

[11] Galvan and Payne, "Implicit Bias as a Cognitive Manifestation of Systemic Racism."

[12] Ibid., 112.

[13] Alice Xiang, "Mirror, Mirror, on the Wall, Who's the Fairest of Them All?" *Dædalus* 153 (1) (Winter 2024): 250–267, https://www.amacad.org/publication/mirror-mirror-wall-whos-fairest-them-all; and Darren Walker, "Deprogramming Implicit Bias: The Case for Public Interest Technology," *Dædalus* 153 (1) (Winter 2024): 268–275, https://www.amacad.org/publication/deprogramming-implicit-bias-case-public-interest-technology.

[14] Morehouse and Banaji, "The Science of Implicit Race Bias," 38; and Xiang, "Mirror, Mirror, on the Wall, Who's the Fairest of Them All?"

[15] Morehouse and Banaji, "The Science of Implicit Race Bias"; Alexandra Kalev and Frank Dobbin, "Retooling Career Systems to Fight Workplace Bias: Evidence from U.S. Corporations," *Dædalus* 153 (1) (Winter 2024): 213–230, https://www.amacad.org/publication/retooling-career-systems-fight-workplace-bias-evidence-us-corporations; and Jack Glaser, "Disrupting the Effects of Implicit Bias: The Case of Discretion and Policing," *Dædalus* 153 (1) (Winter 2024): 151–173, https://www.amacad.org/publication/disrupting-effects-implicit-bias-case-discretion-policing.

[16] Meltzoff and Gilliam, "Young Children & Implicit Racial Biases."

[17] Morehouse and Banaji, "The Science of Implicit Race Bias."

[18] Kalev and Dobbin, "Retooling Career Systems to Fight Workplace Bias."

[19] Wanda A. Sigur and Nicholas M. Donofrio, "Implicit Bias versus Intentional Belief: When Morally Elevated Leadership Drives Transformational Change," *Dædalus* 153 (1) (Winter 2024): 231–249, https://www.amacad.org/publication/implicit-bias-versus-intentional-belief-when-morally-elevated-leadership-drives; and Kalev and Dobbin, "Retooling Career Systems to Fight Workplace Bias."

[20] Kang, "Little Things Matter a Lot"; Anthony G. Greenwald and Thomas Newkirk, "Roles for Implicit Bias Science in Antidiscrimination Law," *Dædalus* 153 (1) (Winter 2024): 174–192, https://www.amacad.org/publication/roles-implicit-bias-science-antidiscrimination-law; and Glaser, "Disrupting the Effects of Implicit Bias."

[21] Hetey, Hamedani, Markus, and Eberhardt, "'When the Cruiser Lights Come On,'" 125.

[22] Hetey, Hamedani, Markus, and Eberhardt, "'When the Cruiser Lights Come On'"; and Marcella Nunez-Smith, "The Case for Data Visibility," *Dædalus* 153 (1) (Winter 2024): 18–20, https://www.amacad.org/publication/case-data-visibility.

Seeing the Unseen

Eric H. Holder, Jr.

Implicit bias has been described as the "tendency for stereotype-confirming thoughts to pass spontaneously through our minds . . . leading to discrimination."[1] Decades of research by social scientists have yielded substantial evidence that measurable, pervasive, and consequential implicit biases exist. This research shows that we can all hold implicit biases, even in the absence of overt prejudice, heartfelt bigotry, or, as we call it in the law, intent. Imagine navigating the world with harmful, stereotype-confirming thoughts deeply embedded in your subconscious framing your day-to-day interactions. Now, imagine navigating the world as someone on the receiving end of implicit bias. Imagine your day-to-day interactions with people who have harmful, stereotype-confirming thoughts about you. Imagine those people are not just passersby on the street, but people you engage with over the course of your lifetime: your doctors, teachers, neighbors, judges, employers and coworkers, local law enforcement officers, and so on.

When you think of implicit bias this way, it is much easier to conceptualize it as a threat to our individual and collective livelihood, including the bedrock principle of equal justice under the law. At its core, the legal profession is based on judgment and discretion. Unfortunately, statistics prove that implicit biases are widespread and have infected both judgment and discretion. Black motorists are almost two times more likely to be searched, despite statistics suggesting they are less likely to be carrying illegal contraband.[2] Yet Black motorists are less likely to be stopped at night, when it is more difficult to discern the race of a motorist.[3] Despite similar rates of drug involvement, Black people are disproportionately incarcerated for drug offenses.[4] As a sad acknowledgment of this reality, Black parents all across America continue to sit down their Black children to explain how they should interact with police if they are ever stopped or confronted in a manner that feels unwarranted.[5]

Studies have shown that partisan actions based on implicit biases may impact the voting process. Minority communities disproportionately have fewer polling locations and older voting machines.[6] A study about the 2012 elections showed that state legislators were less likely to respond to email inquiries regarding voter identification requirements when sent from an account that bore the name of a

https://doi.org/10.1162/daed_a_02045

fictional Latino person versus a fictional white person.[7] Moreover, a survey of the 2008 elections found that the race of both the poll worker and the voter affected the rate at which voters from different groups were asked for identification, but overall, Black and Hispanic voters were asked to show picture ID more often than white voters.[8]

Our founding principles commit us to the ideal of equal justice under the law. These statistics reflect a different reality – that we have not fully realized that principle. Our efforts to overcome implicit bias begin with a commitment to understanding the problem. We must tirelessly study disparities to understand both the source and scope of implicit bias, and ferret out the ill-effects, wherever they may lie. That is why publications like this issue of *Dædalus* are so important. We are unable to address things we do not understand.

We must then commit ourselves to education and solutions. We must show that our institutions take seriously their most solemn responsibility: equality under the law. We must ensure that everyone understands the importance of being aware of, and having strategies to counteract, the unconscious biases we all hold. This is not easy work, and the effects of implicit biases will not disappear quickly. But this work deserves our steadfast attention, as we all hold a stake in the pursuit of a more perfect union.

ABOUT THE AUTHOR

Eric H. Holder, Jr., a Fellow of the American Academy since 2020, is Senior Counsel at Covington & Burling LLP. He served as the 82nd Attorney General of the United States from 2009 to 2015.

ENDNOTES

[1] Keith Payne, Laura Niemi, and John M. Doris, "How to Think about 'Implicit Bias,'" *Scientific American*, March 27, 2018, https://www.scientificamerican.com/article/how-to-think-about-implicit-bias.

[2] Elizabeth Pierson, Camelia Simoiu, Jan Overgoor, et al., "A Large-Scale Analysis of Racial Disparities in Police Stops across the United States," *Nature Human Behaviour* 4 (2020): 736–745, https://doi.org/10.1038/s41562-020-0858-1; and Jordan Bennett, "Research Shows Black Drivers More Likely to Be Stopped by Police," NYU News, May 5, 2020, https://www.nyu.edu/about/news-publications/news/2020/may/black-drivers-more-likely-to-be-stopped-by-police.html.

[3] Ibid.

[4] Jolene Foreman, *From Prohibition to Progress: A Status Report on Marijuana Legalization* (New York: Drug Policy Alliance, 2018), https://drugpolicy.org/news/prohibition-progress-what-new-york-needs-know-about-marijuana-legalization; Nazgol Ghandnoosh, "Black Lives Matter: Eliminating Racial Inequity in the Criminal Justice System" (Washington, D.C.: The Sentencing Project, 2015), https://www.sentencingproject.org/publications/black-lives-matter-eliminating-racial-inequity-in-the-criminal-justice-system; and Ezekiel Edwards, Will Bunting, and Lynda Garcia, *The War on Marijuana in Black and White: Billions of Dollars Wasted on Racially Biased Arrests* (New York: American Civil Liberties Union, 2013), https://www.aclu.org/publications/report-war-marijuana-black-and-white.

[5] Shannon Malone Gonzalez, "Making It Home: An Intersectional Analysis of the Police Talk," *Gender & Society* 33 (3) (2019): 363–386, https://doi.org/10.1177/0891243219828340; Abril Harris and Ndidiamaka Amutah-Onukagha, "Under the Radar: Strategies Used by Black Mothers to Prepare Their Sons for Potential Police Interactions," *Journal of Black Psychology* 45 (6–7) (2019): 439–453, https://doi.org/10.1177/0095798419887069; and Sonia Sotomayor, dissenting opinion, *Utah v. Strieff*, 579 U.S. 232, 136 S. Ct. 2056 (2016).

[6] The Leadership Conference Education Fund, *Democracy Diverted: Polling Place Closures and the Right to Vote* (Washington, D.C.: The Leadership Conference on Civil and Human Rights, 2019), https://civilrights.org/democracy-diverted.

[7] Matthew S. Mendez and Christian R. Grose, "Doubling Down: Inequality in Responsiveness and the Policy Preferences of Elected Officials," *Legislative Studies Quarterly* 43 (3) (2018): 457–491, https://doi.org/10.1111/lsq.12204.

[8] R. Michael Alvarez, Stephen Ansolabehere, Adam Berinsky, et al., *2008 Survey of the Performance of American Elections: Final Report* (Philadelphia: The Pew Charitable Trusts, 2010), https://dspace.mit.edu/bitstream/handle/1721.1/49847/Final%20report20090218.pdf.

The Case for Data Visibility

Marcella Nunez-Smith

Bias is ingrained within the fabric of American society, and as we strive toward healthy communities, we must strive toward equity. To do so, it is essential that we consider not only the most obvious forms of bias, but also the embedded, often unconscious, prejudices that permeate every workplace, institution, and policy. Within the fight for health equity, efforts to counter implicit bias must be ever present. The history of medicine is rife with discrimination, oppression, and exploitation of marginalized populations. This is evident through well-known instances of racism in medicine such as the U.S. Public Health Service Study of Untreated Syphilis in the Negro Male or Henrietta Lacks, but even more so, this remains clear throughout the lived experiences of patients of color who face daily disproportionate discrimination in medical encounters.[1] This history, along with the systemic structures that it intersects, generates contemporary health disparities. To achieve health justice, we must address systemic and embedded biases.

To move the needle on health equity, we cannot only analyze the presence of bias in existing policies; we must also proactively counter ongoing impacts of bias and discrimination. For this reason, the Biden-Harris administration prioritized health equity in the fight against COVID-19. The COVID-19 pandemic underscored and exacerbated deep health disparities in this country. To address this and to ensure equitable access to COVID-19 therapeutics and vaccines, President Biden signed an Executive Order establishing the COVID-19 Health Equity Task Force on his first full day in office.[2] As Chair of the Presidential Task Force and Senior Advisor to the White House COVID-19 Response Team, I worked in partnership with community, academic, government, and industry leaders to advance access and equity in the national response to COVID-19.

Data is imperative to driving public health responses, yet implicit bias pervades our public health and medicine data ecosystems. This includes invisibility and erasure in data, which hide the depth of health inequities in this country and enable the ongoing structural violence perpetuated by health disparities. To achieve health equity, accurate and comprehensive data collection on wide-ranging demographics and social determinants of health – including race and ethnicity – is fundamental.

https://doi.org/10.1162/daed_a_02046

Thus, in combatting COVID-19 health disparities, we were dedicated to collecting data for the hardest hit communities and identifying data sources that would support the execution of equitable access to personal protective equipment, testing, vaccines, and therapies. To accomplish this, the Biden-Harris administration took a multipronged approach, which included assessing the nationwide collection of demographic and socioeconomic variables; expecting that all government entities collect, analyze, and share information on demographic and socioeconomic variables; leveraging the Centers for Disease Control and Prevention's Social Vulnerability Index (SVI) to guide vaccination venue location; and identifying data shortfalls and challenges to better prepare and respond to future pandemics.

By centering partnerships and data equity to address implicit and explicit bias in the COVID-19 response, we were able to change the course of COVID-19 health disparities. For example, in May 2021, at the beginning of the COVID-19 vaccine rollout, only 53 percent of eligible Black Americans had received the first dose of the vaccine compared to 63 percent of white Americans. Through leveraging the SVI, addressing social determinants of health, and centering trustworthy community messengers, the administration intervened and made vaccinations more accessible to communities disproportionately impacted by the pandemic. Coordinated and collective partnerships resulted in historic vaccination parity by January 2022, eliminating racial/ethnic gaps in adult COVID-19 vaccination rates.

To have far-reaching impact, we must confront implicit bias at every level and across every sector. Changemaking also demands an unequivocal focus on marginalized populations. Thus, as we address existing frameworks and develop new ones, we must place historically marginalized and minoritized communities at the forefront and align incentives toward health equity. Only with this intentional consideration can we advance health justice.

ABOUT THE AUTHOR

Marcella Nunez-Smith (MD, MHS) is the Associate Dean for Health Equity Research, C.N.H Long Professor of Internal Medicine, Public Health, and Management, and Director of the Equity Research and Innovation Center at the Yale School of Medicine. An elected member of the National Academy of Medicine, she was the Chair of the Presidential COVID-19 Health Equity Task Force and Senior Advisor to the White House COVID-19 Response Team, Cochair of the Biden-Harris Transition COVID-19 Advisory Board, and Chair of Governor Ned Lamont's Reopen Connecticut Advisory Group Community Committee.

ENDNOTES

[1] For information on the U.S. Public Health Service Study of Untreated Syphilis in the Negro Male, see "About the USPHS Syphilis Study," Tuskegee University, https://www.tuskegee.edu/about-us/centers-of-excellence/bioethics-center/about-the-usphs-syphilis-study (accessed December 22, 2023). For information on Henrietta Lacks, see Denise Grady, "A Lasting Gift to Medicine that Wasn't Really a Gift," *The New York Times*, February 1, 2010, https://www.nytimes.com/2010/02/02/health/02seco.html.

[2] Exec. Order No. 13,995, 86 Fed. Reg. 7193 (Jan. 21, 2021).

The Science of Implicit Race Bias: Evidence from the Implicit Association Test

Kirsten N. Morehouse & Mahzarin R. Banaji

Beginning in the mid-1980s, scientific psychology underwent a revolution – the implicit revolution *– that led to the development of methods to capture implicit bias: attitudes, stereotypes, and identities that operate without full conscious awareness or conscious control. This essay focuses on a single notable thread of discoveries from the Race Attitude Implicit Association Test (RA-IAT) by providing 1) the historical origins of the research, 2) signature and replicated empirical results for construct validation, 3) further validation from research in sociocognitive development, neuroscience, and computer science, 4) new validation from robust association between regional levels of race bias and socially significant outcomes, and 5) evidence for both short- and long-term attitude change. As such, the essay provides the first comprehensive repository of research on implicit race bias using the RA-IAT. Together, the evidence lays bare the hollowness of current-day actions to rectify disadvantage experienced by Black Americans at individual, institutional, and societal levels.*

The science of implicit race bias emerged from a puzzle. By the 1980s, laboratory experiments and surveys revealed clear and noteworthy reductions in expressions of racial animus by White Americans toward Black Americans.[1] But on every dimension that determines life's opportunities and outcomes – housing, employment, education, health care, treatment by law and law enforcement – the presence of widespread racial inequality remained. Further, on surveys asking even slightly indirect questions, such as attitudes toward federal support for racial equality in employment, attitudes appeared to have regressed, with 38 percent support in 1964 dropping to 28 percent in 1996.[2] These inconsistencies demanded an answer from science.

In their search for an explanation, experimental psychologists recalled an interesting *dissociation* or disparity in beliefs recorded decades ago. During his travels through the Jim Crow South, Gunnar Myrdal, a Swedish economist engaged by the Carnegie Corporation to conduct a study on interracial relations in America,

https://doi.org/10.1162/daed_a_02047

encountered an unexpected dilemma. The data from surveys and interviews of White Americans confirmed expected expressions of racism. And yet as Myrdal noted, other sentiments from the very same individuals spoke to their uneasy acknowledgment of a disparity between the cherished national ideal of equality and the history of slavery and the realities of racism, even decades after emancipation. These dissonant cognitions, expressed inside quiet homes and noisy factories, struck Myrdal as distinctive enough to serve as the motif for his classic treatise, *An American Dilemma : The Negro Problem and Modern Democracy*.[3]

Four decades later, psychologists responded to receding levels of "old-fashioned racism" by generating theories of "aversive racism" and measures of "modern racism."[4] These ideas emerged as necessary acknowledgment that although race bias persists, modern racism manifests in more indirect and subtle ways than before. Indeed, experimental data emerging in the 1980s further highlighted the presence of automatic race bias in the minds of honest race egalitarians.[5] With accumulating evidence demonstrating that many judgments and decisions could operate outside conscious awareness or control, social psychologists Anthony G. Greenwald and Mahzarin R. Banaji proposed the idea of *implicit bias* and suggested that a tractable measure of implicit cognition was needed.[6] This essay reports on a thread of the development and discoveries of a singularly important test : the Race Attitude Implicit Association Test (hereafter, RA-IAT), a measure designed to capture differential automatic attitudes, such as associations of "good" and "bad" with White and Black Americans.[7]

In 1967, Martin Luther King Jr. gave the keynote address at the annual meeting of the American Psychological Association (APA), only months before his assassination. He seemed to be aware that his audience of largely White Americans was eager to learn how they could contribute to the success of the civil rights movement. But King's speech clearly conveyed his perspective regarding the responsibility of the APA's scholars and clinicians. If they wished to support the movement, they should simply "'tell it like it is.'"[8] This essay is a response to that call from more than fifty years ago, to emphasize the strength and pervasiveness of anti-Black bias today. We tell it like it is, believing that empirical knowledge production is indeed the responsibility of scientists with expertise in psychological and other sciences. However, the responsibility of addressing challenges to the ideal of racial justice sits squarely at the feet of the nation. In fact, it would be ill-advised to expect scientists – who generally lack knowledge of history, law, policy development, organizational behavior, and the modes of societal change – to be primarily responsible for imagining and constructing paths to social change. By telling it like it is, and remaining focused on the evidence itself, this report can, should the will exist, serve as a foothold to move America toward a solution to racial inequality.

History and Definitions

The science of implicit bias is rooted in experimental psychology. At the core of a particular family of measures is the concept of *mental chronometry*: studying the mind by measuring the time course of human information processing.[9] That is, rather than analyzing participants' responses to a question, the critical unit of measurement is the response latency or the time it takes to react to a stimulus. In the 1970s, researchers conducted the first robust studies testing the automaticity of *semantic memory*. These studies indexed the strength of association between two concepts by using precisely timed stimuli and measuring an individual's response latencies on the order of tens of milliseconds.[10] These procedures were soon adapted to test another important dimension of word meaning: *valence*, that is, the *good-bad* or *pleasant-unpleasant* dimension. Evidence soon emerged that, like semantic meaning, word or concept valence could be automatically extracted by relying on response latencies.[11] Today, this result is received wisdom, and evaluative priming is regarded to be a standard method to measure *automatic attitudes*.[12]

This class of experimental procedures captured the attention of psychologists concerned with the limitation of self-report measures of racism: individuals can withhold their true beliefs in favor of more socially desirable responses. Moreover, even if the desire to speak forthrightly is assured, self-report measures are limited because humans have a desire to present a positive view of themselves, not just to others but even to themselves. Finally, even if such concerns about self and social desirability were removed, a great deal of research had demonstrated that access to mental content and process is vastly limited, making the problem less an issue of motivation and more one of inaccessibility.[13] These considerations, especially the latter, led psychologists to adapt mental chronometry to study automatic or implicit forms of bias. Race was a natural domain for exploration because of the inconsistency between conscious values in aspirational documents like the U.S. Constitution and the history of American racism.

A harbinger of the breakthrough to come appeared in a paper by psychologists John F. Dovidio, Nancy Evans, and Richard Tyler.[14] Diverging notably from previous research methods, these researchers sat their subjects before a computer screen on which the category labels "Black" or "White" appeared. After each of these primes, target words that represented positive and negative stereotypes of these groups (such as ambitious, sensitive, stubborn, lazy) appeared on the screen, and subjects were asked to decide rapidly if each stereotypic word could "ever be true" or was "always false" of the group. The results were clear: participants classified words more quickly when positive words followed "White" and when negative words followed "Black" primes, suggesting that the category White was more positive than Black in participants' implicit cognition. Although this method lacked the components that are characteristic of standard measures of implicit

cognition today (the response task still required deliberation), this study pointed toward the potential of nonreactive measurement of race bias.

Social psychologist Patricia Devine's dissertation experiments hammered a second stake into the ground.[15] She subliminally presented words that captured negative Black stereotypes (in the experimental condition) or neutral words (in the control condition) and then requested evaluations of an ambiguously described person. Remarkably, those who were subliminally exposed to Black stereotypes as primes were more likely to view the ambiguously described person as hostile than those in the control condition. Equally remarkable, the degree of race bias on this more automatic measure of stereotypes was similar *regardless* of consciously reported levels of anti-Black prejudice.

Devine's research demonstrated the first classic dissociation between more deliberate or explicit race attitudes and more automatic or implicit race attitudes, and it prompted a shift in thinking about the nature of race bias. If bias were hidden, even to the person who carried it, that would explain how racial animus could decrease on survey measures while bias embedded in individual minds, institutions, and long-standing societal structures persisted. The two were *dissociated*. From a research standpoint, it was clear that to gain access to race bias in all forms, experimental psychologists would need to develop and sharpen measures of implicit race bias.

Several measures of implicit cognition emerged, among them the Implicit Association Test (IAT).[16] The IAT followed in the tradition of its predecessors by relying on a single fundamental idea: when two things become paired in our experience (for instance, *granny* and *cookies*), evoking one (granny) will automatically activate the other (cookies). In the context of race bias, the speed and accuracy with which we associate concepts like *Black* and *White* with attributes like *good* and *bad* provides an estimate of the strength of their mental association, in this case, an implicit attitude.

Today, decades after the first uses of terms such as *implicit bias*, *implicit attitude*, and *implicit stereotype*, these concepts have permeated scientific and scholarly writing as well as the public's consciousness so effectively that they are rarely accompanied by a definition or explanation.[17] The earliest formal definition of implicit cognition reads: "The signature of implicit cognition is that traces of past experience affect some performance, even though the influential earlier experience is not remembered in the usual sense – that is, it is unavailable to self-report or introspection."[18]A more colloquial definition of implicit bias has emerged as "a form of bias that occurs automatically and unintentionally, that nevertheless affects judgments, decisions, and behaviors."[19]

Both definitions are quite general, and wisely so, to be inclusive of any domain under investigation (such as self-perception, health decisions, and financial decisions). However, despite its generality, the greatest empirical attention has been

devoted to one particular family of biases: those that concern attitudes (valence) and stereotypes (beliefs) about *social groups* (such as by age, gender, sexuality, race, ethnicity, social class, religion, or nationality). Among these, the test that has garnered the greatest scientific and public interest is the race test (as seen in the scientific record and from completion rates of the test online, where the RA-IAT outstrips all other tests in public interest).[20] Unsurprisingly, and for the same reasons, some resistance to the science of implicit race bias has also emerged, but such criticisms remain minor (2 percent of thousands of Google Alerts analyzed include any critical commentary).[21]

Scope of the Essay

Although full-fledged research on implicit social cognition began only in the 1990s, thousands of research articles on implicit bias have since been published. In fact, Google Scholar returns over sixty-five thousand results in response to a query of *implicit bias* as of January 2024. This prolificacy, while notable, renders any complete review of the literature impossible. As such, this essay constrains coverage in four ways. First, we report research on implicit race attitudes, setting aside all other social categories (such as gender, age, sexuality, disability) with a focus on construct validity. Second, we highlight research on *attitudes*, setting aside research on race *stereotypes*. Third, we focus almost entirely on a single method, the IAT, because 1) it is the most widely used measure of implicit bias today (the original report by Greenwald, Debbie McGhee, and Jordan L. K. Schwartz has recorded over seventeen thousand citations on Google Scholar as of January 2024), and 2) the online presence and popularity of the RA-IAT at Project Implicit offer an unparalleled source of data to explore implicit race attitudes.[22] Surprisingly, the signature results from this most popular IAT over the last twenty-five years have not been presented in a single location before. We synthesize them here. Fourth and finally, given the mission of *Dædalus* to explore the frontiers of knowledge on issues of public importance, we prioritize coverage of questions about the *nature* of implicit race bias and its interpretation rather than questions of primarily scientific interest, such as the nature of the psychological *processes* underlying implicit bias, like whether the underlying representation is best viewed as associative or propositional in nature.[23]

With these constraints and opportunities in mind, we introduce 1) streams of research from other sciences, notably cognitive development, neuroscience, and computer science, to provide convergent validation for the RA-IAT data; 2) new research providing predictive validity by demonstrating robust covariation between regional RA-IAT and racial disparities in health care, education, business, and treatment by law enforcement; and 3) evidence demonstrating the RA-IAT's malleability at the individual level (change within one person) and population

level (change within the United States). Together, the data offer confidence in the concept of *implicit race bias* for use in two ways: as a foothold to an effort for broad-based programs and procedures to ensure racial equality, and as the basis for teaching about implicit bias in all educational settings, including schools, colleges, and the workplace.

The Race Attitude IAT: Early Discoveries and Signature Results Providing Validation

Evidence of implicit race bias using the IAT first emerged in the mid-1990s from small-scale, highly controlled experiments administered to college students, as was characteristic of research at that time. These initial experiments were important for benchmarking data that would soon arrive from exponentially larger and more diverse internet-based samples. In 1998, Yale University hosted a test of implicit race attitude, the RA-IAT, among a few other IATs, and the site was immediately bombarded with participants. The RA-IAT was immediately the most popular test, and it remains so twenty-five years later. Today, the amount of research conducted and the diversity of empirical results obtained may appear insurmountable to the general reader. Here, we have created the first repository of the basic discoveries and signature results of the RA-IAT in easy-to-access percentages, histograms, and inferential statistics.

Implicit Social Cognition Terminology and IAT Components

The RA-IAT, following the general IAT procedure, consists of items that appear on a computer screen belonging to a pair of target categories (such as *Black* and *White*) and a pair of target attributes (such as *Good* and *Bad*). At the most basic level, the RA-IAT provides an index of implicit race bias by measuring the relative speed (on the order of milliseconds) it takes participants to sort stimuli when White and Good share a response key (and Black and Bad share a different response key), relative to when Black and Good share a response key (and White and Bad share a different response key).[24] The IAT score is captured by the statistic D, which is a measure of effect size, computed by taking the difference between response latencies in the two critical conditions (that is, Black + Good/White + Bad, and Black + Bad/White + Good) and divided by the standard deviation across all blocks of the test.

Uninitiated readers may wish to take the test at https://implicit.harvard.edu/implicit/selectatest.html. Additionally, in Table 1, we provide descriptions and examples of the core terminology of implicit social cognition and the IAT more generally, even though our focus in this essay will remain on the concept of the attitude.

Table 1
Core Terminology of Implicit Social Cognition Theory and the Implicit Association Test

Term	Description	Labels (examples)	Stimuli (examples)
Concept Category	The concept or category of scientific interest: that is, the target object toward which a measure of attitude or stereotype is sought, such as race, gender, age, sexuality	Black, White, Asian, Latinx (race) Male, Female, Nonbinary (gender) Elderly, Young (age)	Photos/pictures to represent the concept (such as faces of Black and White individuals) Names or other words to represent the concept (such as John or Jane to represent gender) Faces or images to represent age
Attribute Category	The psychological process of scientific interest such as attitude, stereotype, identity; the attribute is the category whose strength of association to the concept category is tested	**Attitude:** Good-Bad, Pleasant-Unpleasant **Stereotype:** Strong-Weak, Smart-Dumb, Honest-Lying **Identity:** Me-Not Me, Me-Other	**Good:** Love, peace, joy **Bad**: Devil, awful, failure **Strong:** Powerful, sturdy, robust **Weak:** Fragile, delicate, frail **Me:** Me, Myself, I, Mine **Not Me**: Not Me, They, Them, Other
Attitude	Evaluative or valence dimension	Good-Bad, Pleasant-Unpleasant, Positive-Negative	See "Attribute Category" row for example stimuli
Stereotype	Beliefs about social groups	Strong-Weak, Smart-Dumb, Honest-Lying	See "Attribute Category" row for example stimuli
Identity	Attitudes and beliefs about oneself	Me-Not Me, Me-Other	See "Attribute Category" row for example stimuli

Source: Descriptions and definitions by the authors.

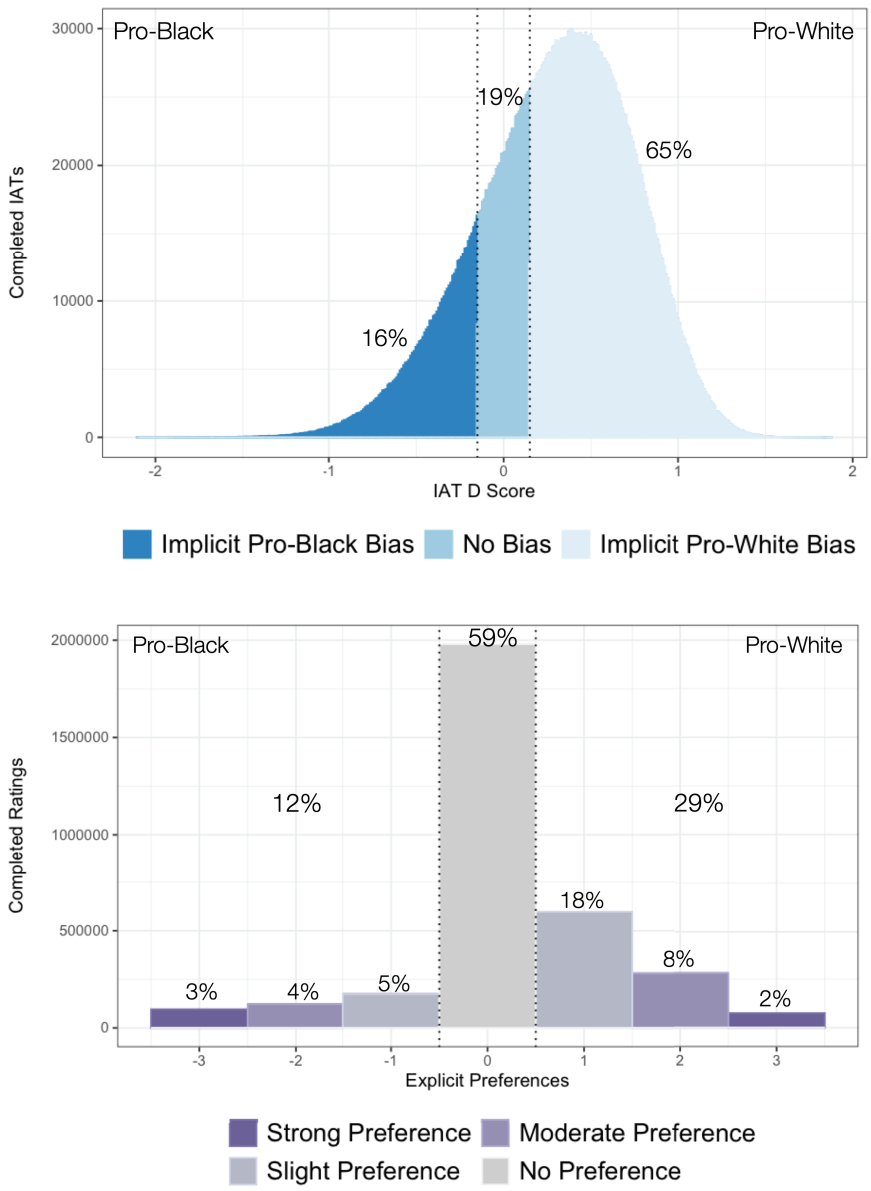
Overall Levels of Explicit and Implicit Race Attitudes and Their Dissociation

An analysis of Project Implicit data from 3.3 million American respondents who completed the RA-IAT across fourteen years (2007–2020) shows robust evidence of implicit race bias: overall, 65 percent of respondents displayed a meaningful association of White with good relative to Black with good ("implicit pro-White bias"), whereas 19 percent of respondents displayed no preference (see Figure 1; for corresponding effect sizes, see Table 2).[25] That is, 2.1 of 3.3 million respondents automatically associated the attribute "Good" (relative to "Bad") more so with White than Black Americans. By contrast, across all fourteen years, only 29 percent of respondents *explicitly* reported a preference for White over Black, and 60 percent of respondents reported equal liking for both groups. As the reader may anticipate, these overall scores are strongly modulated by the social group of the respondent; those data are presented in the next section.

This divergence between mean levels of implicit and explicit race attitudes is striking and bolstered by a dissociation between implicit and explicit race attitudes within a single person. Specifically, modest correlations between implicit and explicit attitudes are typically observed across all participants (for example, $r = 0.30$ [95% CI: 0.308, 0.310]), and even weaker correlations often emerge for Black Americans (see Table 2).[26] Additional support for this dissociation has been derived from latent variable modeling. Unlike variables that can be directly observed or measured (like temperature), latent variables refer to constructs – such as race attitudes – that are inferred indirectly and can possess a degree of measurement error. These latent modeling techniques indicate that implicit and explicit attitudes are related, but distinct. That is, although the latent implicit and explicit attitude variables are correlated ($r = 0.47$), a confirmatory factor analysis suggests that a two-factor solution fits the data better than a single-factor solution with a single latent "attitude" variable.[27] In other words, this technique indicated that implicit and explicit attitudes are related, but psychometrically distinct.

Together, this pattern of data – low levels of explicit race bias but high levels of implicit bias – is considered a key result of implicit intergroup cognition. The data also provide a conceptual replication of Devine's early discovery that implicit race bias can emerge in defiance of stated egalitarian values.[28] However, unlike Devine's work with subliminally presented stimuli, the IAT does not hide its intent; the two racial categories are in full view and the test is announced as one of race bias. Moreover, the IAT components are not shrouded in mystery and completing the task is so simple that even a child can participate. These features contribute to the surprise that often accompanies the IAT: if the task itself is easy, why can I not control my responses?

Figure 1
Distributions of Implicit and Explicit Race Attitudes



IAT D scores range from -2.0 to 2.0, with 0 ± 0.15 serving as the null interval ("Little or No Bias"). Source: Created by the authors using Project Implicit data.

Table 2
Implicit and Explicit Race Attitudes by Participants' Race/Ethnicity

Demographic Subgroup		Implicit		Explicit		
	N	IAT D	Cohen's *d*	Mean	Cohen's *d*	E-I Correlation
Overall	3,325,990	0.29	0.66	0.20	0.19	0.30 [0.308, 0.310]
White	1,881,719	0.36	0.85	0.42	0.51	0.21 [0.211, 0.214]
Asian (East and South)	176,218	0.30	0.70	0.28	0.29	0.27 [0.261, 0.271]
Hispanic	335,780	0.25	0.57	−0.02	−0.02	0.27 [0.275, 0.281]
Multiracial	43,650	0.15	0.34	−0.19	0.62	0.28 [0.268,0.285]
Black	290,837	−0.05	−0.11	−1.07	−0.84	0.17 [0.164, 0.171]

IAT D scores range from −2 to +2, with positive values indicating an implicit pro-White bias. Explicit preferences ranged from −3 ("I strongly prefer African Americans to White Americans") to +3 ("I strongly prefer White Americans to African Americans"). The column "E-I" represents the correlation between IAT D scores and explicit preferences, with 95 percent confidence intervals reported in brackets. Source : Compiled by the authors using Project Implicit data.

Nevertheless, after nearly a century of work based on almost purely explicit measures, these results lay bare the full extent of the challenge we face when confronting the status of race in America today.[29] Recall in Myrdal's interviews during Jim Crow that respondents revealed a disparity between two consciously held beliefs : the American ideal of liberty and equality and America's history of bondage and inequality. In a sense, that conflict is psychologically simple because both cognitions are conscious. By contrast, the dissociation between explicit and implicit race attitudes is especially challenging because implicit attitudes operate largely outside the purview of conscious awareness and control, and therefore may unwittingly produce behaviors that conflict with consciously held values and beliefs.

Explicit and Implicit Race Bias by Racial/Ethnic Group

Among psychology's most ubiquitous results is the demonstration of in-group bias. Irrespective of whether the groups involved are "minimal" (based on a "minimal" preference, such as for the artist Klee over Kandinsky) or real, research has overwhelmingly demonstrated that humans show a preference for their own group relative to the out-group.[30] For example, Japanese Americans and Korean Americans, Yankee and Red Sox fans, and Yale and Harvard students all display clear and symmetric in-group preferences.[31] However, as visualized in Figure 2, the data across White and Black Americans paint a much more complex picture.

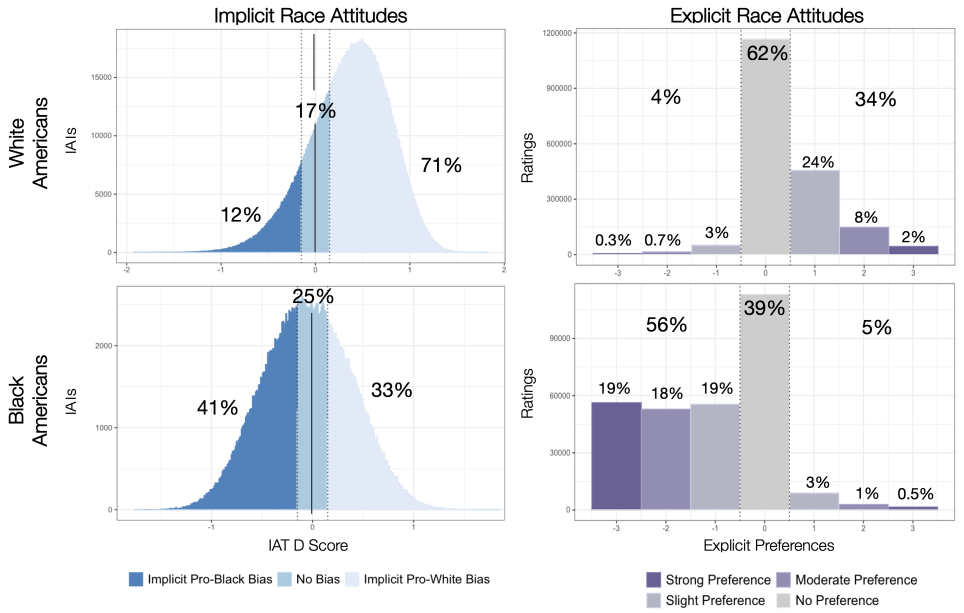
Specifically, 71 percent of White Americans displayed an implicit pro-White bias, whereas only 33 percent of Black Americans displayed an implicit pro-Black bias. These data are in contrast with the robust in-group preferences among Japanese and Korean Americans, Red Sox and Yankee fans, and Yale and Harvard students, in which each group showed an equally robust preference for its own group. This lack of in-group preference among Black Americans is a second signature result and it extends beyond Black Americans to other less advantaged groups. That is, unlike members of socially advantaged groups, who consistently display implicit in-group preferences, members of socially disadvantaged groups typically do not.

On the measure of *explicit* bias, an almost opposite pattern emerges, making these data among the clearest examples of mental dissociation: the lack of consistency between two measures *of the same concept*, within the same mind. Only 34 percent of White Americans displayed an explicit pro-White bias, whereas 56 percent of Black Americans displayed an explicit pro-Black bias. These data highlight the role conscious values play on responses. White Americans, likely being aware of the history of race relations in America, report a far more muted in-group preference. Black Americans, equally likely aware of the history of race relations in America, report an overwhelming in-group preference.

When taken together, the data for White and Black Americans showed a double dissociation. On the one hand, White Americans report little in-group preference on the explicit measure but strong in-group preference on the implicit measure. On the other hand, Black Americans show a strong in-group preference on the explicit measure but no in-group preference on the implicit measure. We regard this result to be sufficiently important that we recommend that it play a role in any discussion of policies to ensure racial equality. Conscious attitudes need not follow such a pattern, but to the extent that attitudes and behavior are driven by both *explicit and implicit* cognition, the balance sheet of intergroup liking shows a striking lack of parity.

Interestingly, when third-party groups are tested (such as Asian Americans taking a White-Black IAT), they consistently show an implicit pro-White bias (see Ta-

Figure 2
Distributions of Implicit and Explicit Race Attitudes for White and Black Americans



IAT D scores range from -2.0 to 2.0, with 0 ± 0.15 serving as the null interval ("Little or No Bias"). Source: Created by the authors using Project Implicit data.

ble 2). That is, rather than associating both out-groups with good equally, third-party respondents display an implicit preference for the socially dominant group. In fact, rivaling the degree of bias among White Americans, 65 percent of Asian Americans and 60 percent of Latinx Americans display an implicit pro-White preference.

Similar patterns also emerge on measures of implicit *stereotyping*. As one example, Morehouse and Banaji, with Keith Maddox, found that White Americans and third-party participants associate human (versus nonhuman attributes like "animal" and "robot") more with their group, whereas nondominant groups (like Black Americans) display no "human = own group" bias.[32] This striking absence of in-group preference in members of disadvantaged groups points to the power of the social standing of groups in society, and has been interpreted to be consistent with system justification tendencies.[33]

Explicit and Implicit Race Bias by Other Demographic Variables

Beyond race/ethnicity, do other demographic variables modulate the strength of implicit race bias? That is, will men and women, liberals and conservatives, or older and younger respondents show different levels of implicit race bias? To test this question, variation across five additional demographic characteristics was examined: religion, level of education, age, gender, and political ideology. Implicit race bias was largely stable across respondents' religious affiliation and level of education. However, differences emerged across age, gender, and political ideology. Implicit pro-White preferences increased with age (each five-year increase translating roughly to a 3 percent increase in IAT D scores), and respondents over age sixty displayed levels of bias that were 15 percent stronger than individuals under age twenty. Further, the incidence of pro-White bias was 20 percent higher among self-identified conservatives relative to self-identified liberals, and 7 percent higher among men relative to women.

These results show how group membership is related to variation in implicit and explicit race attitudes. Later in this essay, we explore another potential determinant of attitude strength – participants' local environment – and the relationship between regional levels of implicit race attitudes and socially significant outcomes (such as lethal use of force by police or health outcomes).

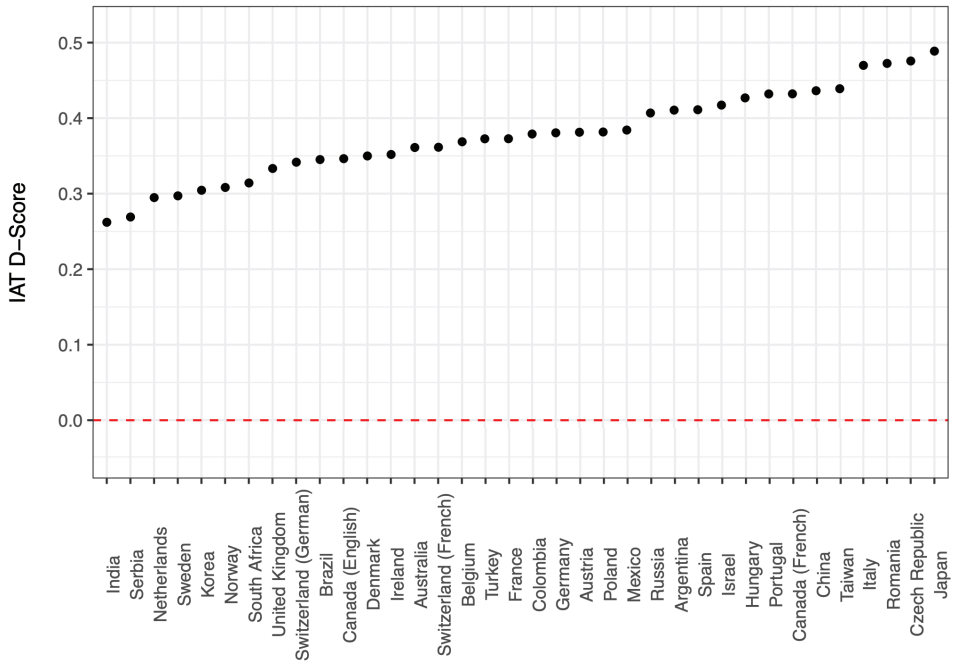
Origins of Implicit Race Bias: Evidence for Developmental Invariance

Over the past twenty years, researchers have gained a new understanding about the surprisingly early precursors of race encoding and race preference in infants and young children. Although far from biological and social maturity, infants and children show evidence of a mind that is already attuned to race but has the capacity to set racial groupings aside, even when attending to other social categories like gender and age, in other situations.[34]

Human groups across the world, as much as they differ by language, culture, preferences, beliefs, and values, are all members of the same species. Is implicit bias a core capacity that unifies us as humans? If we look cross-culturally, a recent analysis of implicit race attitudes from thirty-four countries revealed that an implicit preference for White over Black appears in every country sampled (see Figure 3).[35]

Another way to test whether a particular attitude is fundamental is to observe whether it is present in infants and young children. Our interest here is not in children qua children, but rather in developing minds. Is implicit race bias present even in early stages of cognitive-affective development? The obvious prediction would be that, of course, given the massively different levels of personal experience and knowledge of the culture that children have acquired relative to adults,

Figure 3
Implicit Race Attitudes by Country



Country-level RA-IAT scores expressed in Cohen's *d* effect sizes, with positive effect sizes representing an implicit pro-White bias. For comparison, the average IAT D-Score for the United States for the same period (2009–2019) was 0.30. Source : Adapted from Tessa Charlesworth, Mayan Navon, Yoav Rabinovich, Nicole Lofaro, and Benedek Kurdi, "The Project Implicit International Dataset : Measuring Implicit and Explicit Social Group Attitudes and Stereotypes across 34 Countries (2009–2019)," *Behavioral Research Methods* 55 (3) (2023) : 1413–1440.

implicit race bias should differ based on age. But to the extent that the data show the opposite – similar patterns of implicit race bias in adults and children – we would learn that such biases require little time and experience in a culture to be acquired.

Much has been written about the development of race cognition in infancy.[36] From this work, we know that even infants prefer faces of members of their own group, an effect that likely emerges out of familiarity with their caregivers. For example, three-month-old Ethiopian infants in Ethiopia prefer African over European faces, Ashkenazi babies in Israel prefer European over African faces, and ba-

bies of Ethiopian Jews who have *immigrated* to Israel and have caregivers of both groups show no race preference.[37] Importantly, these preferences are early emerging but not hard-wired; they are absent at birth but present by three months of age.[38] In other words, these data show that the human brain is attuned to features, like race and gender, in the environment that can differentiate between in-group and out-group members.

Work with toddlers has been especially fruitful because the *same method* used to measure implicit race bias in adults could be adapted to measure implicit race bias in children. Specifically, psychologist Andrew Scott Baron and Banaji created a child version of the RA-IAT.[39] Given that children's experiences and knowledge of racial groups vastly differ from adults', the authors expected stark differences in the degree of implicit race bias expressed by children and adults. However, this is not what they found. The surprising result, now replicated many times, is that White six-year-olds, ten-year-olds, and adults show identical levels of implicit race bias.

Notably, and further mirroring the results obtained in adult samples, children's implicit race bias was qualified by social status. By age three, White American children show an in-group preference, whereas Hispanic and Black American children show no in-group preference.[40] This result is remarkable because it teaches us that implicit attitudes are absorbed from the culture and into the minds of even young children. It also challenges the theoretical intuition that implicit attitudes are learned slowly over time. (For further discussion of the development of implicit racial bias, see Andrew N. Meltzoff and Walter S. Gilliam's contribution to this volume.)[41]

Converging Evidence from Neurons and Natural Language

Understanding how the mind works is not for the meek. The Nobel Prize–winning physicist Murray Gell-Mann seemed to understand this when he reputedly said, "Think how hard physics would be if particles could think." Not only are beings who can think the object of our study, but the thinking under consideration is not easily available to their own conscious awareness. As such, building a case for an imperceptible yet consequential bias requires a multipronged, continuous, and iterative process of validation.

There is already deep and broad evidence for the construct validity of the IAT. For example, providing face validity, we know *a priori* that the concept "flower" is more positive than "insect," and the IAT detects this implicit pro-flower preference in most humans.[42] Further evidence can be obtained by studying groups who are known to differ in attitude and observing whether expected differences emerge. Indeed, we have already reported that Black and White Americans show diverging implicit race attitudes, providing additional evidence for construct validity. As a

third route, construct validation has been obtained by demonstrating that findings derived on the IAT are related to (but not redundant with) conceptually similar constructs. Indeed, we have shown that although implicit and explicit race attitudes are modestly correlated, latent variable modeling suggests that a two-factor solution (with "implicit bias" and "explicit bias" as separate latent factors) provided the best fit to the data. In fact, providing discriminant validation, implicit insect-flower attitudes did not hang together with implicit intergroup attitudes.

In the following sections, we will encounter construct validation in several new ways. In particular, we show that methods from other fields (including neuroimaging and word embeddings) also demonstrate evidence of implicit race bias. Moreover, we explore the origins and consequences of implicit race bias to push the engine of construct validity further. Together, these various approaches have not only created a strong foundation for understanding the concept of implicit race bias, but have produced unexpected empirical findings that challenged and refined existing theory.

The Neural Basis of the RA-IAT

When the first pre-IAT measures of implicit attitudes were introduced, little discussion ensued about whether these alien measures should be considered measures of *attitude*.[43] However, when the IAT was introduced, the question of construct validity appeared immediately.[44] It became obvious that measures that directly interrogated the brain, especially those regions that had long been identified as playing a role in emotional learning (such as Pavlovian conditioning), could prove useful if correlations between IAT behavior and brain activation patterns in regions known to be evolved in emotional learning could be observed.

Research with neuroimaging methods like fMRI has long demonstrated that the amygdala, a subcortical brain structure, is involved in the continuous evaluation and integration of sensory information, with a special role for assigning values for valence and intensity.[45] Crucially, neuroscientist Elizabeth A. Phelps and colleagues showed that amygdala activation to Black faces of unknown individuals (relative to White) was significantly correlated with implicit race bias; no such correlation was observed with explicit race bias as measured by the Modern Racism Scale.[46] This suggested that whatever the RA-IAT detects has a core valence component, in line with the idea of "attitudes" as measuring evaluations or the dimension of *positive* and *negative*. A second study suggested that race-based responding is modulated by experience: when the faces of famous and generally liked Black (Denzel Washington) and White (Jerry Seinfeld) faces were used, this activation-implicit bias correlation disappeared. Put differently, this result indicated that familiarity can interrupt this relationship, providing two-pronged convergence.

In the decades that have followed, a plethora of evidence has linked implicit attitudes with neural responses to race-based in-group and out-group faces and more downstream decision-making to test the ability to control default, biased responding.[47] Results of relevance demonstrate that 1) the neural representation of race-based attitudes involve a range of overlapping and interacting brain systems, 2) race-based processing of in-group and out-group faces occurs early in the information-processing sequence starting at one hundred milliseconds upon encountering a face, 3) implicit bias observed in brain activity is malleable and responsive to task demands and context, and 4) individual differences exist in the ability to exert control over biased responses, and this control itself can be initiated without awareness as well as involve both inhibition of unwanted responses and the initiation and application of intentional behavior.[48] Crucially, this last piece of evidence highlights the need for proactive interventions. If bias can creep in, even during early visual processing, then it is unrealistic to expect even well-intentioned individuals to prevent bias from impacting their behavior in the moment. Instead, changes that alter the choice structure and prevent bias from entering the decision-making process are more likely to succeed.

Overall, neuroscientific evidence provided important construct validity for the IAT and its presumed measurement of expressions of value along a good-bad dimension. Moreover, it indicated that implicit race bias converges with multiple levels of information processing from the earliest stages of face detection to judgments of behavior.

Word Embeddings Based on Massive Language Corpora Converge with IAT Data

A long history of research on natural language processing (NLP) coupled with the availability of massive language corpora (such as the Common Crawl and Google Books) have created the opportunity to learn how social groups are represented in language on an unprecedented scale. Specifically, mirroring the logic of the IAT, computer scientist Aylin Caliskan and colleagues used word embeddings – a technique that maps words or phrases to a high-dimensional vector space – to understand the relative associations between targets (such as Black and White people) and attributes (such as Good and Bad).[49] Creating a parallel measure, the Word Embeddings Association Test (WEAT), they performed tests of group-attribute associations in language on a trained dataset of eight hundred and forty billion tokens from the internet. In doing so, they replicated the classic implicit race bias finding: European American names were more likely than African American names to be closer (semantically similar) to pleasant words than to unpleasant words.

These approaches have also enabled researchers to ask questions about human attitudes that are beyond the scope of behavioral tools. Experimental psycholo-

gist Tessa Charlesworth, Caliskan, and Banaji used trained databases of historical texts to demonstrate that attitudinal biases toward racial/ethnic groups have remained stable over the course of two centuries (1800–1999).[50] Moreover, just as neuroimaging data showed convergence between theoretically identified brain regions like the amygdala and the RA-IAT but *not* with explicit race bias, analyses of the biases embedded in language suggest that they are related to IATs but not self-report data.[51] In other words, linguistic patterns represent a reservoir for collectively held or culturally imprinted beliefs.[52]

In fact, recent work indicates that algorithms are even capable of refracting beliefs about racial purity.[53] Specifically, information scientist Robert Wolfe, Caliskan, and Banaji showed that CLIP, an algorithm that relies on both image and text data, has learned the one-drop rule or hypodescent (that is, a legal principle prominent even in the twentieth century that held that a person with just one Black ancestor is to be considered Black).[54] Overall, these findings add to the burgeoning evidence that implicit bias embedded in human minds exists in language and that algorithms trained on these databases will carry, amplify, and even reproduce bias.[55]

Covariation between Regional Implicit Race Bias and Socially Significant Outcomes

A growing number of "audit studies" have demonstrated group-based discrimination in controlled field settings.[56] These studies, typically conducted by economists and sociologists, create highly standardized but naturalistic situations to explore how specific variables (such as race/ethnicity) influence behavior. For example, economist Marianne Bertrand and computation and behavioral scientist Sendhil Mullainathan sent roughly five thousand fictitious résumés to employers in Boston and Chicago.[57] The résumés were identical in all ways except that the applicant's name was either a White- or Black-sounding name. Despite their identical qualifications, résumés with White names received 50 percent more callbacks than résumés with Black names. In another example in the domain of employment, Devah Pager and colleagues demonstrated that, despite having equivalent résumés and being actors trained to respond identically to interview questions, Black applicants were half as likely to receive a callback than White applicants.[58] In fact, in an even more stunning demonstration of race bias, Black applicants were just as likely to receive a callback as White applicants with a felony record. These individual studies mirror a larger trend observed in a meta-analysis : hiring discrimination against African Americans remained stable over a twenty-five-year period (1989–2015).[59]

These audit studies, like the perplexing disconnect between consciously reported prejudice and observed inequalities in society, require an explanation. How is it that the same résumé or qualifications can be evaluated more positively

if they are attributed to a White person? We posit that implicit bias is the most likely explanation. The difficulty was that, until recently, no direct link between measures of implicit bias and large-scale race-based discrimination was available. However, a new line of research, now reaching a substantial number of demonstrations, provides the first persuasive evidence that implicit bias is indeed correlated with racial discrimination on socially significant outcomes (SSB) in domains like employment, health care, education, and law enforcement.[60]

Specifically, a mounting body of research across laboratories and disciplines within the social sciences shows that U.S. regions with stronger implicit race bias (measured by the RA-IAT and stereotype IATs) also have larger Black-White disparities in SSBs. In fact, this research has demonstrated covariation between regional implicit race bias and SSBs in four prominent domains: 1) education (including suspension rates and Black-White gaps in standardized test scores);[61] 2) life and economic opportunity (adoption rates and upward mobility);[62] 3) law enforcement (Black-White disparities in traffic stops and the use of lethal force);[63] and 4) health care (Medicaid spending and Black-White gaps in infant birth weight and preterm births).[64] These studies show that implicit bias, measured at the level of individual minds but aggregated across geographic space, reflects race discrimination that cannot otherwise be explained.

Evidence and Interventions for Implicit Attitude Change: Early Evidence of Malleability

With hindsight, we know that implicit bias is malleable. However, this was not always received knowledge or even expected. In the early years of research on implicit bias using the IAT, many primary investigators believed that implicit bias was intractable.[65] Yet even early work raised the possibility that implicit race attitudes were sensitive to perceivers' motivations, goals, and strategies, as well as contextual manipulations.[66] For example, social psychologist Bernd Wittenbrink and colleagues found that negativity toward Black individuals was lower after watching a movie clip depicting Black Americans in a positive setting (relative to a negative setting).[67] Similarly, social psychologist Brian Lowery and colleagues demonstrated that White Americans displayed lower levels of negativity toward Black individuals in the presence of a Black (rather than White) experimenter.[68]

Extending this work, psychologist Calvin Lai and colleagues conducted an important study exploring the comparative efficacy of seventeen interventions designed to reduce implicit race bias.[69] Although these interventions were roughly five-minutes long and only administered once, eight of the seventeen interventions were effective in reducing implicit race bias. The most effective interventions invoked high self-involvement and/or linked Black people with positivity and White people with negativity.[70] By contrast, interventions that required perspective-

taking, asked participants to consider egalitarian values, or induced a positive emotion were ineffective. When participants' attitudes were tested even a few hours after the intervention, none of the eight previously effective interventions produced a continued reduction in implicit race bias.[71] Of course, this temporary (but not durable) change is to be expected; implicit bias should snap back, rubber band–like, to some stable individual, situational, or broader cultural default. In fact, that single presentations of short interventions can produce *any* change is surprising.

But many "light" interventions, often involving a few counterattitudinal associations or a hypothetical written scenario (a paragraph long) presenting counterattitudinal information, do not show long-term change. To us, the lack of long-term change is hardly surprising given the weakness of the interventions. In fact, in such a case, implementing flimsy interventions and looking for long-term effects is a fool's errand; yet well-intentioned investigators with the hope that a sentence or two should wipe out a lifetime of learning have tried them.

Change at the Societal Level

These laboratory studies provide excellent tests of specific interventions, but they are less equipped to test whether implicit bias has changed over the course of years or decades. As such, the key question of whether long-term change was possible remained. However, recent analyses by Charlesworth and Banaji challenged this idea.[72] Specifically, using time-series modeling, they traced almost three million Americans' implicit race attitudes over the course of fourteen years (2007–2020). Crucially, they found evidence of pervasive change: across all participants, implicit race bias decreased by 26 percent, making it the second fastest changing implicit attitude after sexuality attitudes (anti-gay bias), which saw a dramatic 65 percent reduction during the same period.[73] In fact, if trends continue, implicit race attitudes could first touch neutrality in 2035.

Moreover, this change was not restricted to only certain segments of society (for instance, younger and more liberal participants). Rather, pointing to *widespread societal change*, men and women, older and younger, liberal and conservative, and more- and less-educated participants alike all moved toward neutrality.[74] The only exception was that, unlike White participants, who recorded a 27 percent reduction in implicit bias (IAT D score reduced by 0.11 points), Black participants' implicit attitudes remained relatively stable, only changing 0.03 IAT D score points over the fourteen-year period (see Table 3).

This widespread change is remarkable, especially when one considers that not all implicit biases are changing. For example, implicit anti-elderly, anti-disability, and anti-fat biases remained relatively stable over the fourteen-year period. This change toward some social categories but not others begs an important question: what is the *source* of this change?

Table 3
Change in Implicit Race Attitudes by Participants' Race/Ethnicity

Demographic Subgroup	Start Value (2007)	End Value (2020)	Raw Change	% Change
Overall	0.33	0.24	–0.09	–27
White	0.41	0.30	–0.11	–27
Hispanic	0.29	0.18	–0.11	–38
Asian (East and South)	0.32	0.23	–0.09	–28
Black	–0.09	–0.06	0.03	33

"Start Value" refers to the mean IAT D score recorded in January 2007; "End Value" refers to the mean IAT D score recorded in December 2020. Source: Compiled by the authors using Project Implicit data.

We pose this question because of its relevance to the different claims about how to reduce bias, and where resources earmarked for attitude change should be directed. On the one hand, some researchers and practitioners have criticized a focus on change at the individual level (such as deploying appeals of equality to change individual minds). On the other hand, past interventions targeting structural-level change have not eradicated racial inequalities as expected.[75] In fact, change through laws and acts of Congress, if resisted by individuals, may actually prompt reactance and undo progress.[76]

We noted above that implicit anti-gay bias dropped dramatically (64 percent) between 2007 and 2020. What caused this surprising and especially rapid change? We propose that anti-gay bias may possess unique features that allowed such change. For one, sexuality is more easily concealed than a person's race/ethnicity, gender, age, or weight. But we argue that another explanation warrants further investigation: anti-gay interventions occurred at three levels within the same fourteen-year period.

First, change occurred at the *individual level* as children (and adults of all ages) came out to parents, grandparents, friends, neighbors, and coworkers. Love, already in place, trumped even implicit bias. In other words, the concealable nature

of sexuality forced individuals to reconcile their anti-gay attitudes with their positive feelings toward their loved ones ; this choice architecture was not in place for attitudes about other social groups. Second, change occurred at the *institutional level*. Of course, such change was not adopted everywhere, and some organizations were directly hostile to nonheterosexual employees. However, many institutions, like the U.S. military, enacted policies that affirmed the status of same-sex relationships (such as extending health benefits to same-sex partners) even before the country did. Third, change occurred at the *macro level*. Massachusetts and other states legalized same-sex marriages in the early 2000s, and the Supreme Court of the United States followed suit in 2015. In our estimation, it is rare for interventions at all three levels – individual, institutional, and societal – to occur within a short period of time. To our knowledge, change at all three levels within a short time frame has not eventualized for other social groups.

Implicit race bias exists. Support for its presence is undergirded by evidence from other areas of psychology (cognitive, developmental, neuroscience) as well as other behavioral sciences using quite different methods. New evidence shows that regional implicit bias predicts socially significant outcomes of Black-White disparity along several important dimensions that determine life's opportunities and outcomes. To bring hope, data also reveal that implicit bias is malleable. Overall, these data represent one of many robust streams of scientific evidence available today. Together, they call for a nationwide undertaking for change – at the individual, institutional, and societal levels.

ABOUT THE AUTHORS

Kirsten N. Morehouse is a PhD candidate in psychology at Harvard University. She uses computational and behavioral tools to study when and why humans harbor implicit associations that are in conflict with ground truth data and consciously held beliefs. She has published in such journals as *Proceedings of the National Academy of Sciences*, *Current Research in Ecological and Social Psychology*, and *Journal of Personality and Social Psychology*.

Mahzarin R. Banaji, a Fellow of the American Academy since 2008, is the Richard Clarke Cabot Professor of Social Ethics in the Department of Psychology and the first Carol K. Pforzheimer Professor at the Radcliffe Institute for Advanced Study at Harvard University ; and the George A. and Helen Dunham Cowan Chair in Human Dynamics at the Santa Fe Institute. She is the author of *Blindspot : Hidden Biases of Good People* (with Anthony G. Greenwald, 2013).

ENDNOTES

[1] Howard Schuman, Charlotte Steeh, and Lawrence Bobo, *Racial Attitudes in America : Trends and Interpretations* (Cambridge, Mass. : Harvard University Press, 1985).

[2] Howard Schuman, Charlotte Steeh, Lawrence D. Bobo, and Maria Krysan, *Racial Attitudes in America : Trends and Interpretations*, rev. ed. (Cambridge, Mass. : Harvard University Press, 1997).

[3] Gunnar Myrdal, *An American Dilemma : The Negro Problem and Modern Democracy*, volumes 1 and 2 (Oxford : Harper, 1944).

[4] For aversive racism, see John F. Dovidio and Samuel L. Gaertner, "Prejudice, Discrimination, and Racism : Historical Trends and Contemporary Approaches," in *Prejudice, Discrimination, and Racism*, ed. John F. Dovidio and Samuel L. Gaertner (San Diego : Academic Press, 1986), 1–34. For so-called modern racism, see John B. McConahay, "Modern Racism, Ambivalence, and the Modern Racism Scale," in ibid.

[5] Patricia G. Devine, "Stereotypes and Prejudice : Their Automatic and Controlled Components," *Journal of Personality and Social Psychology* 56 (1989) : 5–18, https://doi.org/10.1037/0022-3514.56.1.5.

[6] Anthony G. Greenwald and Mahzarin R. Banaji, "Implicit Social Cognition : Attitudes, Self-Esteem, and Stereotypes," *Psychological Review* 102 (1) (1995) : 4.

[7] Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz, "Measuring Individual Differences in Implicit Cognition : The Implicit Association Test," *Journal of Personality and Social Psychology* 74 (6) (1998) : 1464–1480, https://doi.org/10.1037/0022-3514.74.6.1464.

[8] "King's Challenge to the Nation's Social Scientists," *The APA Monitor* 30 (1) (1999), https://www.apa.org/topics/equity-diversity-inclusion/martin-luther-king-jr-challenge.

[9] R. Duncan Luce, *Response Times : Their Role in Inferring Elementary Mental Organization* (New York : Oxford University Press, 1986) ; and Michael I. Posner, *Chronometric Explorations of Mind* (Oxford : Lawrence Erlbaum, 1978).

[10] David E. Meyer and Roger W. Schvaneveldt, "Facilitation in Recognizing Pairs of Words : Evidence of a Dependence between Retrieval Operations," *Journal of Experimental Psychology* 90 (1971) : 227–234, https://doi.org/10.1037/h0031564 ; and James H. Neely, "Semantic Priming and Retrieval from Lexical Memory : Roles of Inhibitionless Spreading Activation and Limited-Capacity Attention," *Journal of Experimental Psychology : General* 106 (3) (1977) : 226–254, https://doi.org/10.1037/0096-3445.106.3.226.

[11] Russell H. Fazio, David M. Sanbonmatsu, Martha Powell, and Frank R. Kardes, "On the Automatic Activation of Attitudes," *Journal of Personality and Social Psychology* 50 (1986) : 229–238, https://doi.org/10.1037/0022-3514.50.2.229.

[12] For a fuller treatment of the "implicit revolution," see Anthony G. Greenwald and Mahzarin R. Banaji, "The Implicit Revolution : Reconceiving the Relation between Conscious and Unconscious," *American Psychologist* 72 (9) (2017) : 861–871, https://doi.org/10.1037/amp0000238.

[13] Greenwald and Banaji, "Implicit Social Cognition" ; and Richard E. Nisbett and Timothy D. Wilson, "Telling More than We Can Know : Verbal Reports on Mental Processes," *Psychological Review* 84 (1977) : 231–259, https://doi.org/10.1037/0033-295X.84.3.231.

[14] John F. Dovidio, Nancy Evans, and Richard B. Tyler, "Racial Stereotypes : The Contents of Their Cognitive Representations," *Journal of Experimental Social Psychology* 22 (1) (1986) : 22–37, https ://doi.org/10.1016/0022-1031(86)90039-9.

[15] Devine, "Stereotypes and Prejudice."

[16] For comprehensive reviews of measures of implicit cognition, see Bertram Gawronski and Jan De Houwer, "Implicit Measures in Social and Personality Psychology," in *Handbook of Research Methods in Social and Personality Psychology*, ed. Harry T. Reis and Charles M. Judd (Cambridge : Cambridge University Press, 2014), 283–310 ; Brian A. Nosek, Carlee Beth Hawkins, and Rebecca S. Frazier, "Implicit Social Cognition : From Measures to Mechanisms," *Trends in Cognitive Sciences* 15 (4) (2011) : 152–159, https ://doi.org/10.1016/j .tics.2011.01.005 ; and Bertram Gawronski, "Automaticity and Implicit Measures," *Handbook of Research Methods in Social and Personality Psychology*, ed. Reis and Judd.

[17] Mahzarin R. Banaji, Curtis Hardin, and Alexander J. Rothman, "Implicit Stereotyping in Person Judgment," *Journal of Personality and Social Psychology* 65 (1993) : 272–281, https ://doi.org/10.1037/0022-3514.65.2.272 ; and Greenwald and Banaji, "Implicit Social Cognition."

[18] Greenwald and Banaji, "Implicit Social Cognition," 4–5.

[19] "Implicit Bias," National Institutes of Health, https ://web.archive.org/web/2022071 6115620/https ://diversity.nih.gov/sociocultural-factors/implicit-bias (accessed January 26, 2024).

[20] For an analysis of Google alerts on "implicit bias," see Kirsten N. Morehouse, Swathi Kella, and Mahzarin R. Banaji, "Implicit Bias in the Public Eye : Using Google Alerts to Determine Public Sentiment" (in preparation).

[21] Jennifer L. Howell and Kate A. Ratliff, "Not Your Average Bigot : The Better-than-Average Effect and Defensive Responding to Implicit Association Test Feedback," *British Journal of Social Psychology* 56 (1) (2017) : 125–145, https ://doi.org/10.1111/bjso.12168 ; and Alexander M. Czopp, Margo J. Monteith, and Aimee Y. Mark, "Standing up for a Change : Reducing Bias through Interpersonal Confrontation," *Journal of Personality and Social Psychology* 90 (5) (2006) : 784–803, https ://doi.org/10.1037/0022-3514.90.5.784.

[22] Greenwald, McGhee, and Schwartz, "Measuring Individual Differences in Implicit Cognition." As of May 2023, over thirty million completed IATs have been sampled and over seventy million tests have been at least partially sampled on Project Implicit. See Project Implicit, https ://implicit.harvard.edu (accessed May 1, 2023).

[23] For more on the psychological processes, see Benedek Kurdi, Kirsten N. Morehouse, and Yarrow Dunham, "How Do Explicit and Implicit Evaluations Shift ? A Preregistered Meta-Analysis of the Effects of Co-Occurrence and Relational Information," *Journal of Personality and Social Psychology* 124 (6) (2022), https ://doi.org/10.1037/pspa0000329 ; and Benedek Kurdi and Mahzarin R. Banaji, "Implicit Person Memory : Domain-General and Domain-Specific Processes of Learning and Change," PsyArXiv, October 18, 2021, last edited November 18, 2021, https ://doi.org/10.31234/osf.io/hqnfy.

[24] For a detailed review of the IAT, see Kate A. Ratliff and Colin Tucker Smith, "The Implicit Association Test," *Dædalus* 153 (1) (Winter 2024) : 51–64, https ://www.amacad.org/pub lication/implicit-association-test.

[25] Standard interpretations regard 0 ± 0.15 as the null (no bias) interval. When using any deviation away from zero as the cutoff, 75 percent of respondents displayed an implicit

White + Good/Black + Bad association. Tessa E. S. Charlesworth and Mahzarin R. Banaji, "Patterns of Implicit and Explicit Attitudes: IV. Change and Stability from 2007 to 2020," *Psychological Science* 33 (9) (2022), https://doi.org/10.1177/09567976221084257.

[26] Brian A. Nosek, Frederick L. Smyth, Jeffrey J. Hansen, et al., "Pervasiveness and Correlates of Implicit Attitudes and Stereotypes," *European Review of Social Psychology* 18 (1) (2007): 36–88, https://doi.org/10.1080/10463280701489053.

[27] William A. Cunningham, John B. Nezlek, and Mahzarin R. Banaji, "Implicit and Explicit Ethnocentrism: Revisiting the Ideologies of Prejudice," *Personality and Social Psychology Bulletin* 30 (10) (2004): 1332–1346, https://doi.org/10.1177/0146167204264654.

[28] Devine, "Stereotypes and Prejudice."

[29] Mahzarin R. Banaji and Anthony G. Greenwald, *Blindspot: Hidden Biases of Good People* (New York: Delacorte Press, 2013).

[30] Henri Tajfel, Michael Billig, Robert P. Bundy, and Claude Flament, "Social Categorization and Intergroup Behaviour," *European Journal of Social Psychology* 1 (2) (1971): 149–178, https://doi.org/10.1002/ejsp.2420010202.

[31] Steven A. Lehr, Meghan L. Ferreira, and Mahzarin R. Banaji, "When Outgroup Negativity Trumps Ingroup Positivity: Fans of the Boston Red Sox and New York Yankees Place Greater Value on Rival Losses than Own-Team Gains," *Group Processes & Intergroup Relations* 22 (1) (2019): 26–42, https://doi.org/10.1177/1368430217712834; Kristin A. Lane, Jason P. Mitchell, and Mahzarin R. Banaji, "Me and My Group: Cultural Status Can Disrupt Cognitive Consistency," *Social Cognition* 23 (4) (2005): 353–386, https://doi.org/10.1521/soco.2005.23.4.353; and Greenwald, McGhee, and Schwartz, "Measuring Individual Differences in Implicit Cognition."

[32] Kirsten N. Morehouse, Keith Maddox, and Mahzarin R. Banaji, "All Human Social Groups Are Human, but Some Are More Human than Others: A Comprehensive Investigation of the Implicit Association of 'Human' to U.S. Racial/Ethnic Groups," *Proceedings of the National Academy of Sciences* 120 (22) (2023): e2300995120, https://doi.org/10.1073/pnas.2300995120.

[33] John T. Jost, "A Quarter Century of System Justification Theory: Questions, Answers, Criticisms, and Societal Applications," *British Journal of Social Psychology* 58 (2) (2019): 263–314, https://doi.org/10.1111/bjso.12297; John T. Jost, Mahzarin R. Banaji, and Brian A. Nosek, "A Decade of System Justification Theory: Accumulated Evidence of Conscious and Unconscious Bolstering of the Status Quo," *Political Psychology* 25 (6) (2004): 881–919, https://doi.org/10.1111/j.1467-9221.2004.00402.x; and John T. Jost and Mahzarin R. Banaji, "The Role of Stereotyping in System-Justification and the Production of False Consciousness," *British Journal of Social Psychology* 33 (1) (1994): 1–27, https://doi.org/10.1111/j.2044-8309.1994.tb01008.x.

[34] For a review, see Tessa Charlesworth and Mahzarin R. Banaji, "The Development of Social Group Cognition in Infancy and Childhood," in *The Oxford Handbook of Social Cognition*, 2nd edition, ed. Donal E. Carlston, K. Johnson, and Kurt Hugenberg (Oxford: Oxford University Press, in press).

[35] Tessa Charlesworth, Mayan Navon, Yoav Rabinovich, Nicole Lofaro, and Benedek Kurdi, "The Project Implicit International Dataset: Measuring Implicit and Explicit Social Group Attitudes and Stereotypes Across 34 Countries (2009–2019)," PsyArXiv, December 11, 2021, last edited March 21, 2022, https://doi.org/10.31234/osf.io/sr5qv.

[36] For a review, see Charlesworth and Banaji, "The Development of Social Group Cognition in Infancy and Childhood." See also Talee Ziv and Mahzarin R. Banaji, "Representations of Social Groups in the Early Years of Life," in *The SAGE Handbook of Social Cognition*, ed. Susan Fiske and C. Macrae (London : SAGE Publications, 2012), 372–389, https ://doi.org/10.4135/9781446247631.n19.

[37] Yair Bar-Haim, Talee Ziv, Dominique Lamy, and Richard M. Hodes, "Nature and Nurture in Own-Race Face Processing," *Psychological Science* 17 (2) (2006) : 159–163, https ://doi.org/10.1111/j.1467-9280.2006.01679.x.

[38] David J. Kelly, Paul C. Quinn, Alan M. Slater, et al., "Three-Month-Olds, but Not Newborns, Prefer Own-Race Faces," *Developmental Science* 8 (6) (2005) : F31–F36, https ://doi.org/10.1111/j.1467-7687.2005.0434a.x.

[39] Andrew Scott Baron and Mahzarin R. Banaji, "The Development of Implicit Attitudes : Evidence of Race Evaluations from Ages 6 and 10 and Adulthood," *Psychological Science* 17 (1) (2006) : 53–58, https ://doi.org/10.1111/j.1467-9280.2005.01664.x.

[40] Yarrow Dunham, Andrew Scott Baron, and Mahzarin R. Banaji, "Children and Social Groups : A Developmental Analysis of Implicit Consistency in Hispanic Americans," *Self and Identity* 6 (2–3) (2007) : 238–255, https ://doi.org/10.1080/15298860601115344.

[41] For a further discussion of the development of implicit race bias, see Andrew N. Meltzoff and Walter S. Gilliam, "Young Children & Implicit Racial Biases," *Dædalus* 153 (1) (Winter 2024) : 65–83, https ://www.amacad.org/publication/young-children-implicit-racial-biases.

[42] To take a flower-insect IAT, visit https ://outsmartingimplicitbias.org/module/iat.

[43] Fazio, Sanbonmatsu, Powell, and Kardes, "On the Automatic Activation of Attitudes."

[44] Russell H. Fazio (in a personal communication, May 1, 2023) confirmed the easy acceptance of results from semantic priming methods that demonstrated automatic attitudes. The reason the IAT was held to higher standards is likely because its chosen attitude objects were not nonsocial entities like *clouds* and *pizza* but rather social categories like race, gender, sexuality, and age. It is likely that discovery of bias on these topics was simply less palatable, including to psychologists who were not familiar with the research tradition on implicit memory from which these measures were derived.

[45] Joseph E. LeDoux, "Emotion and the Amygdala," in *The Amygdala : Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction* (New York : Wiley-Liss, 1992), 339–351 ; and Goran Šimić, Mladenka Tkalčić, Vana Vukić, et al., "Understanding Emotions : Origins and Roles of the Amygdala," *Biomolecules* 11 (6) (2021), https ://pubmed.ncbi.nlm.nih.gov/34072960.2023.

[46] Elizabeth A. Phelps, Kevin J. O'Connor, William A. Cunningham, et al., "Performance on Indirect Measures of Race Evaluation Predicts Amygdala Activation," *Journal of Cognitive Neuroscience* 12 (5) (2000) : 729–738, https ://doi.org/10.1162/089892900562552.

[47] For reviews, see David M. Amodio and Mina Cikara, "The Social Neuroscience of Prejudice," *Annual Review of Psychology* 72 (1) (2021) : 439–469, https ://doi.org/10.1146/annurev-psych-010419-050928 ; Inga K. Rösler and David M. Amodio, "Neural Basis of Prejudice and Prejudice Reduction," *Biological Psychiatry : Cognitive Neuroscience and Neuroimaging* 7 (12) (2022) : 1200–1208, https ://doi.org/10.1016/j.bpsc.2022.10.008 ; Jennifer T. Kubota, Mahzarin R. Banaji, and Elizabeth A. Phelps, "The Neuroscience of Race," *Nature Neuroscience* 15 (7) (2012) : 940–948, https ://doi.org/10.1038/nn.3136 ; Pascal Mo-

lenberghs, "The Neuroscience of In-Group Bias," *Neuroscience & Biobehavioral Reviews* 37 (8) (2013): 1530–1536, https://doi.org/10.1016/j.neubiorev.2013.06.002; and Jennifer T. Kubota, "Uncovering Implicit Racial Bias in the Brain: The Past, Present & Future," *Dædalus* 153 (1) (Winter 2024): 84–105, https://www.amacad.org/publication/uncovering-implicit-racial-bias-brain-past-present-future.

[48] Amodio and Cikara, "The Social Neuroscience of Prejudice"; and Rösler and Amodio, "Neural Basis of Prejudice and Prejudice Reduction."

[49] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan, "Semantics Derived Automatically from Language Corpora Contain Human-like Biases," *Science* 356 (6334) (2017): 183–186, https://doi.org/10.1126/science.aal4230.

[50] The top ten traits associated with White (versus Black): critical, polite, hostile, decisive, friendly, diplomatic, understanding, philosophical, able, and belligerent. The top ten traits associated with Black (versus White): earthy, lonely, cruel, sensual, lifeless, deceitful, helpless, rebellious, meek, and lazy. Tessa E. S. Charlesworth, Aylin Caliskan, and Mahzarin R. Banaji, "Historical Representations of Social Groups across 200 Years of Word Embeddings from Google Books," *Proceedings of the National Academy of Sciences* 119 (28) (2022): e2121798119, https://doi.org/10.1073/pnas.2121798119.

[51] Sudeep Bhatia and Lukasz Walasek, "Predicting Implicit Attitudes with Natural Language Data," *Proceedings of the National Academy of Sciences* 120 (25) (2023): e2220726120, https://doi.org/10.1073/pnas.2220726120.

[52] For an exploration of gender biases embedded in internet texts, see Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, et al., "Gender Bias in Word Embeddings: A Comprehensive Analysis of Frequency, Syntax, and Semantics," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (New York: Association for Computing Machinery, 2022), 156–170, https://doi.org/10.1145/3514094.3534162.

[53] Arnold K. Ho, Jim Sidanius, Daniel T. Levin, et al., "Evidence for Hypodescent and Racial Hierarchy in the Categorization and Perception of Biracial Individuals," *Journal of Personality and Social Psychology* 100 (3) (2011): 492–506, https://doi.org/10.1037/a0021562.

[54] Robert Wolfe, Mahzarin R. Banaji, and Aylin Caliskan, "Evidence for Hypodescent in Visual Semantic AI," in *2022 ACM Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2022), 1293–1304, https://doi.org/10.1145/3531146.3533185.

[55] See also Darren Walker, "Deprogramming Implicit Bias: The Case for Public Interest Technology," *Dædalus* 153 (1) (Winter 2024): 268–275, https://www.amacad.org/publication/deprogramming-implicit-bias-case-public-interest-technology; and Alice Xiang, "Mirror, Mirror, on the Wall, Who's the Fairest of Them All?" *Dædalus* 153 (1) (Winter 2024): 250–267, https://www.amacad.org/publication/mirror-mirror-wall-whos-fairest-them-all.

[56] For reviews, see S. Michael Gaddis, "An Introduction to Audit Studies in the Social Sciences," in *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*, ed. S. Michael Gaddis (Cham: Springer International Publishing, 2018), 3–44, https://doi.org/10.1007/978-3-319-71153-9_1; and S. Michael Gaddis, "Understanding the 'How' and 'Why' Aspects of Racial/Ethnic Discrimination: A Multi-Method Approach to Audit Studies," SSRN, July 25, 2019, https://doi.org/10.2139/ssrn.3426846.

[57] Marianne Bertrand and Sendhil Mullainathan, "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review* 94 (4) (2004): 991–1013, https://doi.org/10.1257/0002828042002561.

[58] Devah Pager, Bruce Western, and Bart Bonikowski, "Discrimination in a Low-Wage Labor Market: A Field Experiment," *American Sociological Review* 74 (5) (2009): 777–799, https://doi.org/10.1177/000312240907400505.

[59] Lincoln Quillian, Devah Pager, Ole Hexel, and Arnfinn H. Midtbøen, "Meta-Analysis of Field Experiments Shows No Change in Racial Discrimination in Hiring over Time," *Proceedings of the National Academy of Sciences* 114 (41) (2017): 10870–10875, https://doi.org/10.1073/pnas.1706255114.

[60] For a review, see Tessa E.S. Charlesworth and Mahzarin R. Banaji, "The Relationship of Implicit Social Cognition and Discriminatory Behavior," prepublication chapter to appear in *Handbook of Economics of Discrimination and Affirmative Action*, ed. Ashwini Deshpande, https://tessaescharlesworth.files.wordpress.com/2021/05/charlesworth_econ-handbook_final.pdf. For a discussion of the practical significance of these relationships, see Jerry Kang, "Little Things Matter a Lot: The Significance of Implicit Bias, Practically & Legally," *Dædalus* 153 (1) (Winter 2024): 193–212, https://www.amacad.org/publication/little-things-matter-lot-significance-implicit-bias-practically-legally; and Manuel J. Galvan and B. Keith Payne, "Implicit Bias as a Cognitive Manifestation of Systemic Racism," *Dædalus* 153 (1) (Winter 2024): 106–122, https://www.amacad.org/publication/implicit-bias-cognitive-manifestation-systemic-racism.

[61] Travis Riddle and Stacey Sinclair, "Racial Disparities in School-Based Disciplinary Actions Are Associated with County-Level Rates of Racial Bias," *Proceedings of the National Academy of Sciences* 116 (17) (2019): 8255–8260, https://doi.org/10.1073/pnas.1808307116; and Mark J. Chin, David M. Quinn, Tasminda K. Dhaliwal, and Virginia S. Lovison, "Bias in the Air: A Nationwide Exploration of Teachers' Implicit Racial Attitudes, Aggregate Bias, and Student Outcomes," *Educational Researcher* 49 (8) (2020): 566–578, https://doi.org/10.3102/0013189X20937240.

[62] Sarah Beth Bell, Rachel Farr, Eugene Ofosuc, et al., "Implicit Bias Predicts Less Willingness and Less Frequent Adoption of Black Children More than Explicit Bias," *The Journal of Social Psychology* 163 (4) (2023): 554–565, https://doi.org/10.1080/00224545.2021.1975619; and Raj Chetty, Nathaniel Hendren, Maggie R. Jones, and Sonya R. Porter, "Race and Economic Opportunity in the United States: An Intergenerational Perspective," *The Quarterly Journal of Economics* 135 (2) (2020): 711–783, https://doi.org/10.1093/qje/qjz042.

[63] B. Keith Payne, Heidi A. Vuletich, and Jazmin L. Brown-Iannuzzi, "Historical Roots of Implicit Bias in Slavery," *Proceedings of the National Academy of Sciences* 116 (24) (2019): 11693–11698, https://doi.org/10.1073/pnas.1818816116; and Eric Hehman, Jessica K. Flake, and Jimmy Calanchini, "Disproportionate Use of Lethal Force in Policing Is Associated With Regional Racial Biases of Residents," *Social Psychological and Personality Science* 9 (4) (2018): 393–401, https://doi.org/10.1177/1948550617711229.

[64] Jordan B. Leitner, Eric Hehman, and Lonnie R. Snowden, "States Higher in Racial Bias Spend Less on Disabled Medicaid Enrollees," *Social Science & Medicine* 208 (2018): 150–157, https://doi.org/10.1016/j.socscimed.2018.01.013; and Jacob Orchard and Joseph Price, "County-Level Racial Prejudice and the Black-White Gap in Infant Health Outcomes," *Social Science & Medicine* 181 (2017): 191–198, https://doi.org/10.1016/j.socscimed.2017.03.036.

[65] Mahzarin R. Banaji, "The Opposite of a Great Truth Is Also True: Homage of Koan #7," in *Perspectivism in Social Psychology: The Yin and Yang of Scientific Progress* (Washington, D.C.: American Psychological Association, 2004), 127–140, https://doi.org/10.1037/10750-010.

[66] For a review, see Irene V. Blair, "The Malleability of Automatic Stereotypes and Prejudice," *Personality and Social Psychology Review* 6 (3) (2002): 242–261, https://doi.org/10.1207/S15327957PSPR0603_8.

[67] Bernd Wittenbrink, Charles M. Judd, and Bernadette Park, "Evaluative versus Conceptual Judgments in Automatic Stereotyping and Prejudice," *Journal of Experimental Social Psychology* 37 (3) (2001): 244–252, https://doi.org/10.1006/jesp.2000.1456.

[68] Brian S. Lowery, Curtis D. Hardin, and Stacey Sinclair, "Social Influence Effects on Automatic Racial Prejudice," *Journal of Personality and Social Psychology* 81 (2001): 842–855, https://doi.org/10.1037/0022-3514.81.5.842.

[69] Calvin K. Lai, Maddalena Marini, Steven A. Lehr, et al., "Reducing Implicit Racial Preferences: I. A Comparative Investigation of 17 Interventions," *Journal of Experimental Psychology: General* 143 (4) (2014): 1765–1785, https://doi.org/10.1037/a0036260.

[70] Past research suggests that exposure to only positive Black figures may be less effective at changing implicit racial attitudes than exposure to both positive Black and negative White exemplars. Jennifer A. Joy-Gaba and Brian A. Nosek, "The Surprisingly Limited Malleability of Implicit Racial Evaluations," *Social Psychology* 41 (3) (2010): 137–146, https://doi.org/10.1027/1864-9335/a000020.

[71] Calvin K. Lai, Allison L. Skinner, Erin Cooley, et al., "Reducing Implicit Racial Preferences: II. Intervention Effectiveness across Time," *Journal of Experimental Psychology: General* 145 (8) (2016): 1001–1016, https://doi.org/10.1037/xge0000179.

[72] Tessa E. S. Charlesworth and Mahzarin R. Banaji, "Patterns of Implicit and Explicit Attitudes: I. Long-Term Change and Stability From 2007 to 2016," *Psychological Science* 30 (2) (2019): 174–192, https://doi.org/10.1177/0956797618813087; Tessa E. S. Charlesworth and Mahzarin R. Banaji, "Patterns of Implicit and Explicit Attitudes: IV. Change and Stability From 2007 to 2020," *Psychological Science* 33 (9) (2022), https://doi.org/10.1177/09567976221084257; Tessa E. S. Charlesworth and Mahzarin R. Banaji, "Patterns of Implicit and Explicit Attitudes II. Long-Term Change and Stability, Regardless of Group Membership," *American Psychologist* 76 (6) (2021): 851–869, https://doi.org/10.1037/amp0000810; and Tessa E. S. Charlesworth and Mahzarin R. Banaji, "Patterns of Implicit and Explicit Stereotypes III: Long-Term Change in Gender Stereotypes," *Social Psychological and Personality Science* 13 (1) (2022): 14–26, https://doi.org/10.1177/1948550620988425.

[73] Explicit race attitudes recorded a 98 percent reduction, shifting from a "'slight' preference for White Americans over Black Americans" to neutrality in the span of fifteen years, and making it the fastest changing explicit bias. Charlesworth and Banaji, "Patterns of Implicit and Explicit Attitudes II."

[74] Ibid.

[75] *Brown v. Board of Education of Topeka*, 347 U.S. 483 (1954), https://www.oyez.org/cases/1940-1955/347us483. See also Alexandra Kalev and Frank Dobbin, "Retooling Career Systems to Fight Workplace Bias: Evidence from U.S. Corporations," *Dædalus* 153 (1) (Winter 2024): 213–230, https://www.amacad.org/publication/retooling-career-systems-fight-workplace-bias-evidence-us-corporations.

[76] For a discussion of why legislation is often inadequate, see Wanda A. Sigur and Nicholas M. Donofrio, "Implicit Bias versus Intentional Belief : When Morally Elevated Leadership Drives Transformational Change," *Dædalus* 153 (1) (Winter 2024) : 231–249, https ://www.amacad.org/publication/implicit-bias-versus-intentional-belief-when-morally-elevated-leadership-drives.

The Implicit Association Test

Kate A. Ratliff & Colin Tucker Smith

Among the general public and behavioral scientists alike, the Implicit Association Test (IAT) is the best known and most widely used tool for demonstrating implicit bias: the unintentional impact of social group information on behavior. More than forty million IATs have been completed at the Project Implicit research website. These public datasets are the most comprehensive documentation of IAT and self-reported bias scores in existence. In this essay, we describe the IAT procedure, summarize key findings using the IAT to document the pervasiveness and correlates of implicit bias, and discuss various ways to interpret IAT scores. We also highlight the most common uses of the IAT. Finally, we discuss unanswered questions and future directions for the IAT specifically, and implicit bias research more generally.

MON____
PAN____
SHE____

Fill in the blanks to complete the words above. What did you come up with? Imagine that before responding to these word stems, you were casually exposed to a list of animal names. Research shows that, in that case, you would be more likely to complete the stems with Monkey, Panda, and Sheep than Monday, Pancake, and Sheet. This residual effect of prior learning can occur even if you are unable to recall the animal word list when asked. This example illustrates implicit memory.[1] Although never directly instructed to use previous information, people's responses indicate a residual effect of what they have learned previously.

In 1995, psychologists Anthony G. Greenwald and Mahzarin R. Banaji introduced the idea of *implicit attitudes*, arguing that the processes underlying implicit memory effects can also apply in the social world.[2] In the same way that traces of experience with word lists can influence word stem completions, traces of experiences can also influence evaluations of social groups – even when we are unable to verbally report on those evaluations. Shortly after Greenwald and Banaji first wrote on implicit attitudes, Greenwald published the Implicit Association Test (IAT) as a measure of performance of these implicit social cognitions, including implicit attitudes (evaluations of groups), implicit self-esteem (attitudes toward oneself), and implicit stereotypes (beliefs about traits that are characteristic of a group).[3]

https://doi.org/10.1162/daed_a_02048

In this essay, we describe the IAT procedure, summarize key findings using the IAT, and discuss various ways to interpret IAT scores. We also highlight the most common uses of the IAT. Finally, we discuss unanswered questions and future directions for the IAT specifically, and implicit bias research more generally.

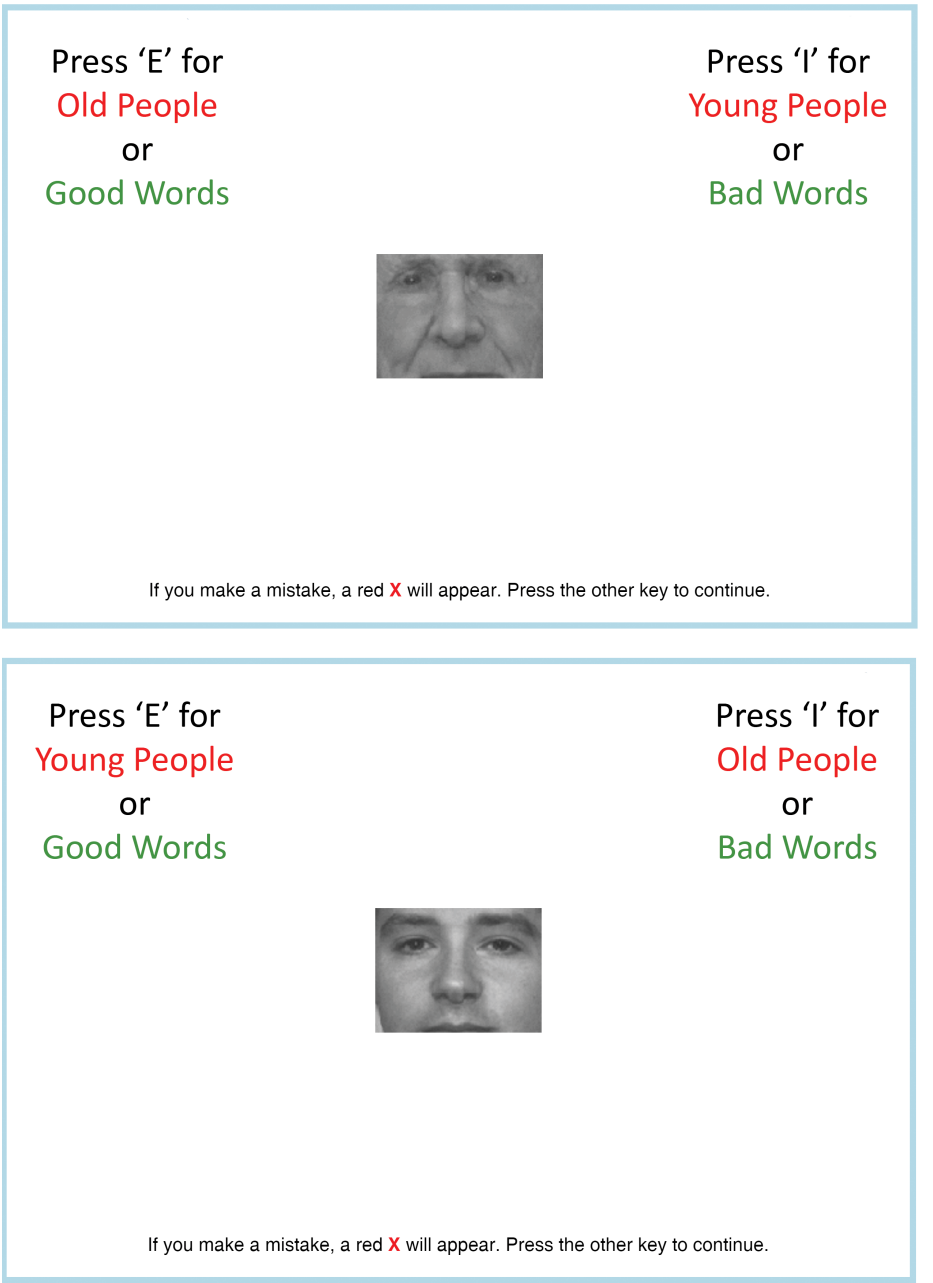
The idea behind the IAT is quite simple: people perform tasks better when a response relies on stronger mental links compared to when a response relies on weaker mental links. Because the IAT is a procedure, not a discrete measure, and researchers vary the features of the task depending on their preferences, there is no single IAT. However, most IATs follow the same general format; let us walk through the age-attitudes version of the task.

Participants in the IAT are tasked with sorting words or pictures into categories as quickly and accurately as possible. There are two key blocks of trials within the IAT in which two categories share the same response (such as a key on a computer keyboard, a square block on a touch device). In the block of trials pictured in Figure 1, if an elderly face appears *or* positive words appear, you would press the "E" key. If a young-adult face *or* a negative word appears, you would press the "I" key. You would first complete a set of trials sorting words and pictures in this way. And then the categories switch so the young-adult faces and positive words share the same response key, and older-adult faces and negative words share the same response key, and you would go through the process again with the updated pairings.

All the while, the computer is recording how long it takes for you to make a correct response on each trial. An IAT score reflects the standardized difference in average response time between the two sorting conditions. If someone completes the task faster when young people and positive words share the same response key, and old people and negative words share the same response key – as in the bottom picture in Figure 1 – their IAT score would reflect an implicit bias favoring young people over old people. If they complete the task faster when old people and positive words share the same response key and young people and negative words share the same response key – as in the top picture – their IAT score would reflect an implicit bias favoring old people over young people.[4]

In 2003, Greenwald and Banaji, together with psychologist Brian Nosek, incorporated Project Implicit, a nonprofit organization with a public education mission and an international research collaboration between behavioral scientists interested in implicit social cognition. The core feature of Project Implicit is a demonstration website, set up in the model of an interactive exhibit at a science museum, where visitors can complete an IAT on a topic of their choice. As of late 2023, more than eighty million study sessions have been launched and more than forty million IATs completed at the Project Implicit website – an IAT every twenty-one seconds.[5] In addition, there is an uncounted multitude of people who

Figure 1
Sample Screens from the Age-Attitudes Implicit Association Test



Source: Age ("Young-Old") Implicit Association Test at Harvard University. See Project Implicit, https://implicit.harvard.edu/implicit/takeatest.html (accessed November 27, 2023).

have interacted with the IAT in classroom settings or as part of an educational session at their place of work.

Over the past twenty-five years, we have learned a lot about implicit bias as measured by the IAT. Greenwald and colleagues' paper introducing the IAT has been cited more than sixteen thousand times since 1998. Across the forty million IATs completed at the Project Implicit website, IAT scores reflect a moderate to strong bias for systematically advantaged groups over systemically disadvantaged or minoritized groups. As seen in Figure 2, there is a clear pattern in favor of straight people (relative to gay people), thin people (relative to fat people), abled people (relative to disabled people), White people (relative to Black people), cisgender people (relative to transgender people), and young people (relative to old people). Notably, people self-report these same biases, but the strength of these biases are considerably weaker.

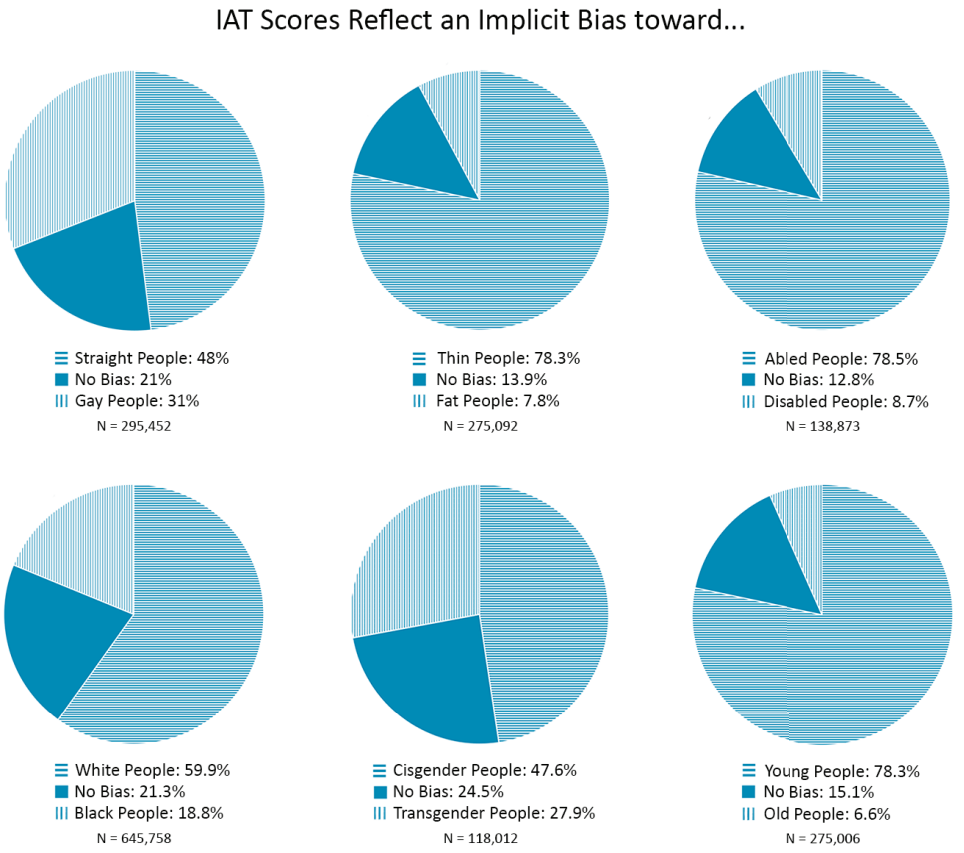
A notable limitation of the IAT, like most other implicit measures, is that it assesses evaluations based on only one clear identity or social group at a time. In real life, of course, people have multiple identities and these identities intersect. In other words, people belong to age and racial and gender groups, and these identities intersect to produce different patterns of experiences, both for the target and perceiver. People's identities in real life are often also far more ambiguous than the stimuli used in implicit measures of bias.

In addition to the direction and strength of an IAT score (that is, which group it favors and whether we describe it as slight, moderate, or strong), we can also think about the pervasiveness of IAT-measured implicit bias by looking at the percentages of respondents on each task whose IAT score indicates a bias favoring one group over another. For example, approximately 67 percent of visitors to the Project Implicit website have an IAT score indicating some degree of implicit bias toward White people (relative to Black people). And we see similar patterns of IAT scores on tasks indicating an implicit bias toward thin people (relative to fat people), abled people (relative to disabled people), straight people (relative to gay people), young people (relative to old people), and cisgender people (relative to transgender people).

Overall, there are few individual variables that consistently relate to IAT scores. Meta-analytically across all the tasks at the Project Implicit site that are about social groups, we see essentially no relationship between IAT scores and education, religiosity, or age, and we see small relationships between IAT scores and prior IATs completed, political orientation, and gender. There are two factors that correlate fairly substantially with IAT scores. One is self-reported attitudes. People who report having more bias also have more biased performance on the IAT. The other factor that matters consistently across almost every task is relevant group membership.

A much higher percentage of heterosexual participants than gay, lesbian, and bisexual participants have an IAT score that reflects bias in favor of straight people:

Figure 2
Proportion of Biases Favoring Dominant over Marginalized Groups in the Implicit Association Test

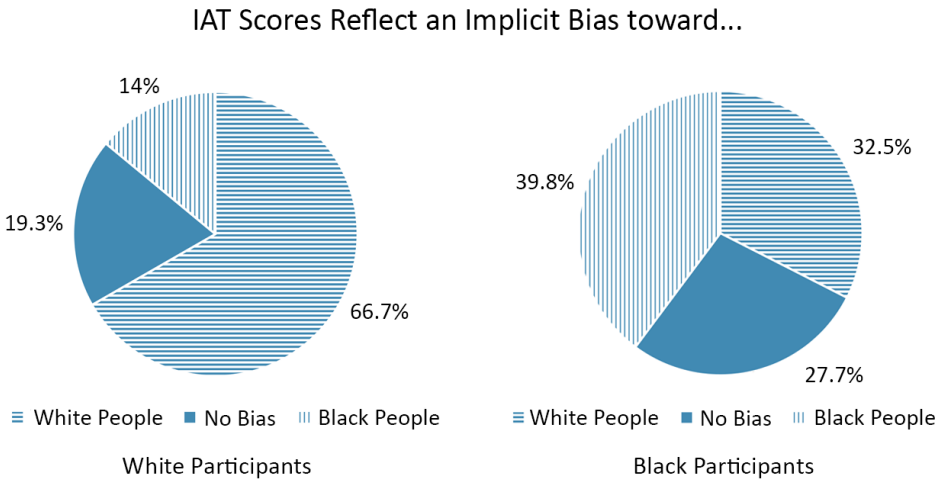


Source: Authors' compilation of data collected at Project Implicit in 2022. Project Implicit, https://osf.io/y9hiq (accessed December 7, 2023).

62 percent compared to 27 percent. Similarly, a higher percentage of White participants than Black participants have an IAT score reflecting an implicit bias toward White people relative to Black people: 73 percent compared to 41 percent. That said, it is not trivial that 41 percent of Black participants have an IAT score reflecting an implicit bias in favor of White people (Figure 3).

Another opportunity that this accumulated data set of IAT scores affords researchers is the ability to track whether levels of implicit bias have changed over

Figure 3
Proportion of Biases Favoring White People over Black People in the Implicit Association Test for White and Black Participants



Source: Authors' compilation of data collected at Project Implicit in 2022. Project Implicit, https://osf.io/y9hiq (accessed December 7, 2023).

time. Banaji and psychologist Tessa Charlesworth summarized patterns of change among 7.1 million data points collected between 2007 and 2020.[6] They found that IAT scores evidencing preferences for young people (relative to old people), abled people (relative to disabled people), and fat people (relative to thin people) have remained fairly stable over time, but preferences for lighter skin (relative to darker skin), White people (relative to Black people), and straight people (relative to gay people) have all decreased in magnitude (that is, shifted toward neutrality over time). This rate of reduction is particularly remarkable for the latter task. Bias favoring straight people (relative to gay people) was reduced by 65 percent across the thirteen-year period sampled. It is also worth noting that these rates of change are happening more quickly for some people than for others. For example, younger people and political liberals showed a larger decrease in implicit anti-gay bias and implicit anti-Black bias than did older people and political conservatives. To be clear, those decreases are evident in all groups, but they are happening faster among some people than others.[7]

Another approach to looking at the influence of time on IAT scores is to compare average IAT scores in some time frame before and after a particular event. For example, the IAT-measured preference for White people (relative to Black people)

in the United States is greater when the economy is worse, and the preference for thin people (relative to fat people) was higher shortly after twenty different highly publicized fat-shaming statements made by celebrities.[8] In addition, the bias on the IAT favoring straight people (relative to gay people) decreased at the state level with implementation of same-sex marriage legalization.[9] In sum, it is clear that IAT scores change slowly over time and also respond to temporary fluctuations in current events.

When drawing so many conclusions based on one data source, it is important to point out that visitors to the Project Implicit website are certainly not representative of the population from which they are drawn. That said, in terms of sheer numbers, the number of data points in the Project Implicit sample is bigger than the total combined population of eighteen U.S. states. It is certainly the largest database of IAT scores in existence and probably the largest for self-reported biases as well. There is also growing evidence that data from Project Implicit samples perform similarly to those collected from nationally representative samples.[10] Thus, because of the scale of IAT data available, it can provide a reasonably good inference about societal-level trends that can complement traditional self-report surveys such as those collected by Gallup or Pew Research Center that rely on random – though generally still not representative – sampling.

You may have noticed that, so far, we have described and discussed IAT *scores*. The data make clear that IAT scores suggest strong and pervasive biases favoring dominant, societally privileged groups over those that are marginalized and minoritized. But how should we think about what IAT scores are, and what implicit bias is?

One of the central tasks of the behavioral sciences is developing procedures and measures to serve as a proxy for psychological constructs. With traditional self-report measures of psychological constructs, this can be straightforward. For example, the ten-item Rosenberg Self-Esteem Scale asks people the extent to which they agree with items like "On the whole, I am satisfied with myself" and "I have a positive attitude toward myself."[11] This type of instrument is high in face validity; in other words, the measurement procedure makes logical sense as a way to assess the construct of interest. The IAT, however, is not as high in face validity. There is quite a leap between the procedure – sorting words and pictures into categories – and what the test purports to measure – evaluations of social groups. Thus, to demonstrate that the IAT can in fact measure evaluations of social groups, we need to look to other kinds of validity. For example, the IAT relates to other measures of evaluations (convergent validity), it does not relate to measures it should be different from (discriminant validity), and it varies based on one's own group memberships, as discussed previously, in ways that make sense (known groups validity).[12] This could be a lengthy discussion, but in sum, the ma-

jority of researchers agree that enough validity evidence has accrued to conclude that the IAT does, in fact, serve as a valid and reliable way to assess individual differences in evaluations of and stereotypes about social groups, though perhaps with a bit more noise than self-report measures.[13]

But let us return to our original questions in this section: what are IAT scores and what is implicit bias? Even after twenty-five years of research, these are still under vigorous debate, with some arguing that the implicitness construct should be done away with altogether due to its ambiguity and lack of precision, or because it offers little above and beyond self-report measures.[14] While we disagree with this conclusion, the value of the implicitness construct *is* one of the most important questions in this line of research, and it is worth summarizing a few of the different ways that scholars think about implicit bias.[15]

The earliest and probably still most common idea is that implicit biases reflect some kind of latent mental construct – a hidden force inside of people's minds – that cannot be directly observed. In this view, implicit biases are something people "have," as in 60 percent of U.S. participants *have* an implicit bias favoring cisgender people over transgender people. In their 1995 paper introducing implicit cognition, Greenwald and Banaji defined implicit attitudes as "introspectively unidentified (or inaccurately identified) traces of past experience that mediate responses."[16] The interpretation of this definition (though perhaps not the intention) is that implicit biases are outside of conscious awareness and inaccessible to introspection. The field's reliance on this definition for more than a decade is likely how *unconscious bias* and *implicit bias* came to be used synonymously. In line with this interpretation, the Project Implicit website defined implicit attitudes and stereotypes for many years as those that people are "unwilling or unable to report."

It has become clear, however, that people do have at least some awareness of their biases, as evidenced by stronger correlations between IAT scores and self-report under particular conditions and by the fact that people are at least somewhat able to predict their IAT scores.[17] It is increasingly obvious that defining implicit bias as an evaluation that is entirely outside of conscious awareness would functionally eradicate the construct, as we currently have no measures that can meet the burden of proof of producing effects that are entirely outside of conscious awareness.[18]

We have argued that if we must distinguish between whether an effect is implicit or explicit bias, (un)consciousness is not the best factor by which to do so because awareness: 1) is complex and multifaceted, 2) is nearly impossible to prove, and 3) ignores the importance of an actor's intentions.[19] Instead, we argue that the key feature of the IAT that distinguishes it from the biases that people self-report is *automaticity*. Psychologists Agnes Moors and Jan De Houwer conceptualize automaticity as a process that influences task performance (that is, behavior in a way

that has one or more of the following features: unintentional, goal-independent, autonomous, unconscious, efficient, and/or fast).[20] Of the particular features of automaticity, intentionality (whether or not one has control over the startup of a process) and control (whether or not one can override a process once started) are highly relevant to distinguishing between implicit and explicit bias.[21]

A vexing problem for the latent mental construct approach to implicit bias is that scores on the IAT and other implicit measures demonstrate group-based preferences that are quite large but are also somewhat unstable. In other words, the same person's score is likely to differ over time, which is not consistent with the idea of deeply ingrained, overlearned unconscious preferences. In response, recent models propose that intergroup attitudes are better understood as group-level constructs. For example, the prejudice-in-places model posits that *places* can be characterized as biased to the extent that they create predictable, systematic inequalities through formal (for example, laws) and informal (for example, norms) mechanisms that disadvantage some groups relative to others.[22] Variations in these regional inequalities then differentially inform individual-level intergroup attitudes. While the prejudice-in-places model does not distinguish between implicit and explicit intergroup attitudes, the "bias of crowds" model takes a similar approach, but focuses on implicit attitudes. It proposes that implicit attitudes across a group of people reflect rather than cause systemic biases. This perspective also assumes that implicit bias reflects what comes to mind most easily at the time, and that measures like the IAT reflect situations more than people. Biases appear stable to the extent that they reflect systemically biased social structures, but they can fluctuate depending on one's current context. The interpretation of this approach is that IAT scores are much better measures of biases held by places than biases held within minds.[23] Or, less radically, that the biases that exist within minds are critically impacted by physical environments.

Support for geographic, intergroup bias comes primarily through research using publicly available data from Project Implicit that aggregate individual IAT scores at some geographic unit (for example, county-level race bias) and then correlate those scores with another indicator that is also aggregated within the same unit, like racial disparities in school discipline, test scores, and police stops.[24] Notably, these county-level differences are not random. History casts a long shadow. For example, IAT scores demonstrating anti-Black bias among White people are higher today in counties and states that were more dependent on the labor of enslaved Black people in 1860, suggesting that historical factors create structural inequalities that are transmitted generationally and that lead to implicit biases favoring White people.[25]

The idea that something as important as racial bias exists in places more so than in people can be a disorienting idea for many of us born and raised within

cultures that predominantly treat places and spaces as neutral and passive while prioritizing the importance of individual actors and their internal states and motivations. In general, when most of us think about a concept like sexism, we think about people (like misogynists). We are unlikely to think about spaces causing people to be sexist. Most researchers have a similar bent. Relatedly, the idea that IAT scores reflect context and history is a radical departure from earlier conceptualizations of implicit bias in two ways, by 1) considering inequality and discrimination as a cause, rather than a consequence, of implicit bias, and 2) implying that countering implicit bias may be accomplished more effectively through changing the environments in which we live rather than changing the individuals who live within those environments.

De Houwer provides a compelling argument that rejects the framing of IAT scores as necessarily reflecting implicit, hidden mental biases that reside inside of minds, and instead conceptualizes performance on measures like the IAT as instances of implicitly biased behavior.[26] The IAT provides an example of how a behavior – the ability to categorize words and pictures – can be influenced by social group cues even when people do not have the intention to be influenced by those cues. Biased responses on more real-world kinds of tasks, like hiring behavior or performance evaluation, can evidence implicit bias even without measures like the IAT that are supposed to assess some kind of mediating attitude or belief. There are two key benefits to this approach. First, a functional approach allows researchers to circumvent the perplexing situation of using the same name ("implicit") for both construct and measure. Second, given that the problem of bias is a behavioral problem, it makes sense to define bias in behavioral terms.

Defining IAT performance as an instance of implicitly biased behavior does not render the results described previously about the pervasiveness of IAT scores favoring privileged groups any less meaningful, nor does it invalidate the idea that performance on the IAT may reflect situations, history, and context more than personal attitudes. Instead, this view positions the IAT as an observable form of bias. This framing requires researchers to explain observable biases rather than engaging in interminable (and potentially intractable) debates about unobservable, theorized mental constructs. For example, it is an observable phenomenon that most participants find it easier to pair bad words with faces of old people than with faces of young people. From there, without mention of underlying processes, we can ask questions such as: Why might they do that? What might that mean? Might some people do that more than others? Can we make people stop doing that?

Before concluding, it is worth discussing the promises and pitfalls of using the IAT as a pre-post measure (testing individuals at different points in time to show

change) to test the efficacy of interventions. For example, imagine an organization assesses the biases of its human resources (HR) team using a gender stereotyping IAT, provides its employees with some kind of training program, and then administers the IAT again, finding a reduction in the IAT score. Success, right? Not necessarily. While it may be reasonable and desirable in some situations to examine bias reduction in this way, there are two important caveats to note. First, research shows that IAT scores tend to move toward zero from one test session to the next, without anything in particular happening in between. Thus, it is critical that anyone using the IAT to assess bias reduction includes a control condition to ensure that the intervention has decreased IAT-measured bias more than it would have decreased anyway. Second, when assessing bias reduction using the IAT (or any measure of group-based bias), it is important to clarify that the bias itself is the construct of interest. Returning to the example of the HR team training, we would encourage this team to consider what the training itself was about and then assess *that*. For example, if the training was about fair interviewing practices, the organization could assess the extent to which HR teams implemented such practices. If the training was about ways to decrease disparities in salary, the organization could assess disparities after a year.

It is difficult to predict what the future holds for the IAT. Citation counts continue to increase year over year, and use of the measure continues to expand into increasingly diverse areas of scholarship. It has been evaluated as rigorously as any psychological measure, and has largely stood up to scrutiny. Further, the concept of "implicit bias" has leapt the walls of the academic journals where it has taken on a life of its own. But ideas ebb and flow, and the way behavioral scientists conceptualize implicit bias has changed dramatically over the last decade, with bias no longer being seen exclusively as a product of individual minds, but instead potentially a product of places. Further, the way that racism and biases exert their power evolves across time, and it is unclear how central implicit forms of bias will be to future versions. We continue to argue about the best ways to define implicit bias in the current time, as evidenced by a recent issue of *Psychological Inquiry* dedicated to the topic.[27] And, as mentioned previously, still others argue that researchers should do away with the term "implicit" altogether.[28] But in doing so, we would lose something important: a language to talk about the indisputable fact that, regardless of where they come from, people have ingrained prejudices and stereotypes that influence how they see and interpret the world. In our view, implicit bias is ordinary, it is rooted in culture, and it is pervasive, and we will continue to need measures like the IAT to document and quantify these biases.

ABOUT THE AUTHORS

Kate A. Ratliff is Associate Professor of Psychology at the University of Florida and past Executive Director at Project Implicit. She has published in such journals as *Journal of Personality and Social Psychology, Psychological Science,* and *Journal of Experimental Psychology: Applied*.

Colin Tucker Smith is Associate Professor of Psychology at the University of Florida. He serves on the Scientific Advisory Board at Project Implicit and has published in such journals as *Journal of Experimental Social Psychology* and *Personality and Social Psychology Bulletin*.

ENDNOTES

[1] Mary S. Weldon, Henry L. Roediger, and Bradford H. Challis, "The Properties of Retrieval Cues Constrain the Picture Superiority Effect," *Memory & Cognition* 17 (1) (1989): 95–105, https://doi.org/10.3758/BF03199561; and Daniel L. Schacter, "Implicit Memory: History and Current Status," *Journal of Experimental Psychology: Learning, Memory, and Cognition* 13 (3) (1987): 501–518, https://doi.org/10.1037/0278-7393.13.3.501.

[2] Anthony G. Greenwald and Mahzarin R. Banaji, "Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes," *Psychological Review* 102 (1) (1995): 4–27, https://doi.org/10.1037/0033-295X.102.1.4.

[3] Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz, "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test," *Journal of Personality and Social Psychology* 74 (6) (1998): 1464–1480, https://doi.org/10.1037/0022-3514.74.6.1464.

[4] For more details regarding IAT procedure and scoring, see Anthony G. Greenwald, Brian A. Nosek, and Mahzarin R. Banaji, "Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm," *Journal of Personality and Social Psychology* 85 (2) (2003): 197–216, https://doi.org/10.1037/0022-3514.85.2.197.

[5] For more information about Project Implicit, see Kate A. Ratliff and Colin Tucker Smith, "Lessons from Two Decades with Project Implicit," in *The Cambridge Handbook of Implicit Bias and Racism*, ed. Jon A. Krosnick, Tobias H. Stark, and Amanda L. Scott (Cambridge: Cambridge University Press, 2023).

[6] Tessa E. S. Charlesworth and Mahzarin R. Banaji, "Patterns of Implicit and Explicit Stereotypes III: Long-Term Change in Gender Stereotypes," *Social Psychological and Personality Science* 13 (1) (2022): 14–26, https://doi.org/10.1177/1948550620988425.

[7] Tessa E. S. Charlesworth and Mahzarin R. Banaji, "Patterns of Implicit and Explicit Attitudes: I. Long-Term Change and Stability from 2007 to 2016," *Psychological Science* 30 (2) (2019): 174–192, https://doi.org/10.1177/0956797618813087; and Charlesworth and Banaji, "Patterns of Implicit and Explicit Stereotypes III."

[8] Emily C. Bianchi, Erika V. Hall, and Sarah Lee, "Reexamining the Link Between Economic Downturns and Racial Antipathy: Evidence that Prejudice Against Blacks Rises During Recessions," *Psychological Science* 29 (10) (2018): 1584–1597, https://doi.org/10.1177/0956797618777214; and Amanda Ravary, Mark W. Baldwin, and Jennifer A. Bartz, "Shaping the Body Politic: Mass Media Fat-Shaming Affects Implicit Anti-Fat Attitudes,"

Personality and Social Psychology Bulletin 45 (11) (2019): 1580–1589, https://doi.org/10.1177/0146167219838550.

[9] Eugene K. Ofosu, Michelle K. Chambers, Jacqueline M. Chen, and Eric Hehman, "Same-Sex Marriage Legalization Associated with Reduced Implicit and Explicit Antigay Bias," *Proceedings of the National Academy of Sciences* 116 (18) (2019): 8846–8851, https://doi.org/10.1073/pnas.1806000116.

[10] Ibid; and Eric Hehman, Jimmy Calanchini, Jessica K. Flake, and Jordan B. Leitner, "Establishing Construct Validity Evidence for Regional Measures of Explicit and Implicit Racial Bias," *Journal of Experimental Psychology: General* 148 (6) (2019): 1022–1040, https://doi.org/10.1037/xge0000623.

[11] Morris Rosenberg, *Society and the Adolescent Self-Image* (Princeton, N.J.: Princeton University Press, 1965).

[12] Anthony G. Greenwald and Calvin K. Lai, "Implicit Social Cognition," *Annual Review of Psychology* 71 (2020): 419–445, https://doi.org/10.1146/annurev-psych-010419-050837; and Anthony G. Greenwald, Miguel Brendl, Huajian Cai, et al., "Best Research Practices for Using the Implicit Association Test," *Behavior Research Methods* 54 (3) (2022): 1161–1180, https://doi.org/10.3758/s13428-021-01624-3.

[13] Paul Connor and Ellen R. K. Evers, "The Bias of Individuals (in Crowds): Why Implicit Bias is Probably a Noisily Measured Individual-Level Construct," *Perspectives on Psychological Science* 15 (6) (2020): 1329–1345, https://doi.org/10.1177/1745691620931492.

[14] Olivier Corneille and Mandy Hütter, "Implicit? What Do You Mean? A Comprehensive Review of the Delusive Implicitness Construct in Attitude Research," *Personality and Social Psychology Review* 24 (3) (2020): 212–232, https://doi.org/10.1177/1088868320911325; and Ulrich Schimmack, "The Implicit Association Test: A Method in Search of a Construct," *Perspectives on Psychological Science* 16 (2) (2021): 396–414, https://doi.org/10.1177/1745691619863798. See also Benedek Kurdi, Kate A. Ratliff, and William A. Cunningham, "Can the Implicit Association Test Serve as a Valid Measure of Automatic Cognition? A Response to Schimmack (2021)," *Perspectives on Psychological Science* 16 (2) (2021): 422–434.

[15] Kate A. Ratliff and Colin Tucker Smith, "Implicit Bias as Automatic Behavior," *Psychological Inquiry* 33 (3) (2022): 213–218, https://doi.org/10.1080/1047840X.2022.2106764; and Bertram Gawronski, Alison Ledgerwood, and Paul W. Eastwick, "Reflections on the Difference Between Implicit Bias and Bias on Implicit Measures," *Psychological Inquiry* 33 (3) (2022): 219–231, https://doi.org/10.1080/1047840X.2022.2115729.

[16] Greenwald and Banaji, "Implicit Social Cognition," 5.

[17] Kate A. Ranganath, Colin Tucker Smith, and Brian A. Nosek, "Distinguishing Automatic and Controlled Components of Attitudes from Direct and Indirect Measurement Method," *Journal of Experimental Social Psychology* 44 (11) (2008): 386–396, https://doi.org/10.1016/j.jesp.2006.12.008; Adam Hahn, Charles M. Judd, Holen K. Hirsh, and Irene V. Blair, "Awareness of Implicit Attitudes," *Journal of Experimental Psychology: General* 143 (3) (2014): 1369–1392, https://doi.org/10.1037/a0035028; and Adam Morris and Benedek Kurdi, "Awareness of Implicit Attitudes: Large-Scale Investigations of Mechanism and Scope," *Journal of Experimental Psychology: General* 152 (12) (2023): 3311–3343, https://doi.org/10.1037/xge0001464.

[18] To be clear, the idea that people harbor internalized prejudices and stereotypes that influence how they see and interpret the world is not dependent on the existence or fitness of any particular measure. Put another way, implicit bias will still exist without the IAT.

[19] Ratliff and Smith, "Implicit Bias"; John A. Bargh, "The Four Horsemen of Automaticity: Awareness, Intention, Efficiency, and Control," in *Handbook of Social Cognition: Basic Processes; Applications*, ed. Robert S. Wyer Jr. and Thomas K. Srull (Mahwah, N.J.: Lawrence Erlbaum Associates, Inc., 1994), 1–40; Adam Hahn and Alexandra Goedderz, "Trait-Unconsciousness, State-Unconsciousness, Preconsciousness, and Social Miscalibration in the Context of Implicit Evaluation," *Social Cognition* 38 (S) (2020): 115–134, https://doi.org/10.1521/soco.2020.38.supp.s115; and Richard E. Nisbett and Timothy D. Wilson, "Telling More Than We Can Know: Verbal Reports on Mental Processes," *Psychological Review* 84 (3) (1977): 231–259, https://doi.org/10.1037/0033-295X.84.3.231.

[20] Agnes Moors and Jan De Houwer, "Automaticity: A Theoretical and Conceptual Analysis," *Psychological Bulletin* 132 (2) (2006): 297–326, https://doi.org/10.1037/0033-2909.132.2.297; and Jan De Houwer and Agnes Moors, "How to Define and Examine the Implicitness of Implicit Measures," in *Implicit Measures and Attitudes*, ed. Bernd Wittenbrink and Norbert Schwarz (New York: The Guilford Press, 2007), 179–194.

[21] Bargh, "The Four Horsemen of Automaticity."

[22] Mary C. Murphy, Kathryn M. Kroeper, and Elise M. Ozier, "Prejudiced Places: How Contexts Shape Inequality and How Policy Can Change Them," *Policy Insights from the Behavioral and Brain Sciences* 5 (1) (2018): 66–74, https://doi.org/10.1177/2372732217748671.

[23] B. Keith Payne, Heidi A. Vuletich, and Kristjen B. Lundberg, "The Bias of Crowds: How Implicit Bias Bridges Personal and Systemic Prejudice," *Psychological Inquiry* 28 (4) (2017): 233–248, https://doi.org/10.1080/1047840X.2017.1335568; and Manuel J. Galvan and B. Keith Payne, "Implicit Bias as a Cognitive Manifestation of Systemic Racism," *Dædalus* 153 (1) (Winter 2024): 106–122, https://www.amacad.org/publication/implicit-bias-as-cognitive-manifestation-systemic-racism.

[24] Travis Riddle and Stacey Sinclair, "Racial Disparities in School-Based Disciplinary Actions Are Associated with County-Level Rates of Racial Bias," *Proceedings of the National Academy of Sciences* 116 (17) (2019): 8255–8260, https://doi.org/10.1073/pnas.1808307116; Mark J. Chin, David M. Quinn, Tasminda K. Dhaliwal, et al., "Bias in the Air: A Nationwide Exploration of Teachers' Implicit Racial Attitudes, Aggregate Bias, and Student Outcomes," *Educational Researcher* 49 (8) (2020): 566–578, https://doi.org/10.3102/0013189X20937240; and Marleen Stelter, Iniobong Essien, Carsten Sander, and Juliane Degner, "Racial Bias in Police Traffic Stops: White Residents' County-Level Prejudice and Stereotypes Are Related to Disproportionate Stopping of Black Drivers," *Psychological Science* 33 (4) (2022): 483–496, https://doi.org/10.1177/09567976211051272.

[25] B. Keith Payne, Heidi A. Vuletich, and Jazmin L. Brown-Iannuzzi, "Historical Roots of Implicit Bias in Slavery," *Proceedings of the National Academy of Sciences* 116 (24) (2019): 11693–11698, https://doi.org/10.1073/pnas.1818816116.

[26] Jan De Houwer, "Implicit Bias Is Behavior: A Functional-Cognitive Perspective on Implicit Bias," *Perspectives on Psychological Science* 14 (5) (2019): 835–840, https://doi.org/10.1177/1745691619855638.

[27] Bertram Gawronski, Alison Ledgerwood, and Paul W. Eastwick, "Implicit Bias ≠ Bias on Implicit Measures," *Psychological Inquiry* 33 (3) (2022): 139–155, https://doi.org/10.1080/1047840X.2022.2106750.

[28] For comparison, see Corneille and Hütter, "Implicit? What Do You Mean?"; and Greenwald and Lai, "Implicit Social Cognition."

Young Children & Implicit Racial Biases

Andrew N. Meltzoff & Walter S. Gilliam

Children are not born harboring racial biases, but they are born learning. Young children, even infants, learn from the "mere observation" of other people's behavior. Nonverbal signals of racial biases are abundant in children's everyday social environments. Studies show that preschool children acquire social group biases when they observe other people's social interactions and nonverbal behaviors. These new findings have implications for child development and educational equity. Even before kindergarten, racial biases are caught even when not explicitly taught, suggesting the need for practical actions for parents, teachers, and others concerned about the transmission of racial bias across generations.

Children are not born with racial biases. Yet children have acquired racial biases before they enter first grade. To construct more complete theories about implicit biases – including sources, consequences, and remedies – we need to understand better how children experience and acquire these biases. How do social group biases held by adults and embodied in societal structures become part of the mental framework of the child?

We would have a ready-made answer to this academic question if parents or schools intentionally provided lessons in racism. In other words, it would be easy to explain young children's racial biases if parents explicitly taught children to reject people of a different race, or if teachers purposely taught racism in the same way they teach reading, through a step-by-step deliberate process, but this is not common. Moreover, the acquisition of racial biases does not follow the principles of classical learning theories for children, such as Skinnerian learning: Psychologists rarely see parents or teachers explicitly reward children for racist words or conduct. Quite the opposite. In the United States today, one commonly encounters parents and teachers who explicitly discourage the expression of racial biases. Despite this, kids acquire these biases, and do so at a very young age.

For developmental psychologists, this presents a puzzle about how infants and young children so readily and effortlessly acquire behaviors, norms, and values from parents and the cultural milieu. Young children, even infants, are no longer believed to be blank slates who primarily learn through Skinnerian reinforcement. Nor are they thought of as Piagetian problem-solvers who construct a conception of the world through independent discovery devoid of social input. Although ef-

https://doi.org/10.1162/daed_a_02049

forts to research the social mechanisms that influence young children's acquisition of racial biases date back to at least the 1920s, only recent experiments have systematically focused on the mechanisms by which these biases are transmitted from adults to young children.[1]

In this essay, we bring modern developmental psychology ideas into the discussion of how young children acquire implicit biases, and we explore the psychological mechanisms underlying the intergenerational transmission of such biases. These mechanisms do not map onto Skinnerian reinforcement, Piagetian constructivism, or deliberate didactic instruction in school. Rather, we describe how children's observational social learning plays a key role in the earliest emergence of implicit biases in children.

Undoubtedly, the most well-known and societally impactful study showing the young age by which children display racial biases was conducted by psychologists Mamie Phipps Clark and Kenneth Clark.[2] Their "doll study" was influential in the 1954 U.S. Supreme Court ruling that school segregation was unconstitutional in *Brown v. Board of Education*. The participants were 253 Black children between three and seven years old who were presented with four dolls that were identical, except that two had white skin and blond hair and the other two had brown skin and black hair. Ninety percent of the children identified the white-skinned, blond-haired doll as being white and the brown-skinned, black-haired doll as being Black. The children were then asked a series of questions regarding their preferences about the dolls and which doll looked most like them. More positive characteristics and preferences were ascribed to the dolls they identified as white, while negative characteristics and rejection were ascribed to the dolls they identified as Black.

Beyond rejection and negative ascriptions, children also display racial biases in judging the amount of pain experienced by others. Psychologist Rebecca A. Dore and colleagues conducted a study in which children were instructed to rate the amount of pain they might feel across ten different events, such as biting their tongue or hitting their head (the children were five, seven, or ten years old, all living in a predominantly white U.S. community, and 90 percent were identified by the parents as being white).[3] The children were then presented with pictures of children matched to their own gender and similar in age, one being a Black child and the other a white child, and they were instructed to rate the pain these children might feel if the same thing happened to them. By seven years of age, children were demonstrating a weak racial bias that the Black child would feel less pain, and by ten years old, the bias was strong and reliable.

These studies and more modern ones with additional experimental controls show that children rapidly acquire racial biases by a young age, including anti-Black biases, and that these biases are linked to the assumptions they make, the de-

gree of empathy they express, and their preferences as to with whom they would prefer to play. What these studies do not show is how young children pick up these biases in the first place and why they are so readily learned.

Children learn more in the first five years of life than at any other equivalent period in development. Adults are wiser than children, and adolescents more introspective about their place in the world, but infants and preschool children are more rapid learners across many domains. For example, infants are born capable of learning any language, and through social interaction with others they quickly become specialized in the language spoken by their cultural group.[4] The answer to the perennial question, "What do children know, and when do they know it?" is that they know more and learn earlier than predicted by classic psychological theories.[5] We will extend this newer developmental framework to implicit biases. No child is born with racial biases, but they acquire them rapidly, often despite parents' efforts to the contrary. How can this be? It seems that racial biases are caught even if not explicitly taught.

A crucial component of how children "catch" racial biases comes from young children's ability for observational social learning and imitation. Although all animals learn, human children are unique in the animal kingdom in their tendency to learn mannerisms, skills, social practices, and values simply by observing the nonverbal behavioral patterns of other people. Social learning and imitation have evolutionary value because they allow human children to learn by "proxy," by watching what others do and reenacting what they see. This fundamental mode of learning about the social and physical world has an underlying basis in the brain and is measurable prior to spoken language.[6] We begin by providing examples of imitation in human infancy because describing this powerful, nonverbal capacity sheds light on how it becomes a channel for the transmission of bias in later childhood.

One study with twelve-month-old infants demonstrated that a twenty-second encounter with a stranger doing something novel with a toy was enough to induce learning.[7] In this laboratory experiment, the infants sat on their mothers' laps and the mothers were blindfolded so that they could not influence the infants' behaviors. After seeing the brief demonstration with the toy (but not being allowed to touch it), the infants and mothers were sent home. The infants were tested after a delay of either one day, one week, or one month. After the delay, some infants were brought back to the same laboratory room, while others went to a completely new environment (to assess generalization across both time and setting). The infants were given the same object they had observed and their behavior was videorecorded. Remarkably, the infants imitated the actions they had seen, even those retested after a one-month delay and in a different room than the original one. The inference is that preverbal infants learn socially from the mere obser-

vation of other people's behaviors and can generalize across space (context) and time (delay). Even though infants at this age are too young to understand what we say, their brains are tuned to remember and imitate what they see us do.

Infants learn not only from what other people do, but also from what they intended to do. In one study, infants watched as an adult tried to pull apart a dumbbell-shaped object, but the adult's hands "accidentally" slipped off the ends so it did not come apart. The adult tried several times in different ways but did not manage to succeed. The infants observed this pattern of actions but were not allowed to handle the toys themselves. There could be no Skinnerian shaping or Piagetian hands-on discovery experience; there was only observation of the adult's behavioral patterns.[8] When the infants were presented with the objects, they did not duplicate what the adult actually did (hands slipping off), but instead reenacted the unspoken goal of the adult's actions. The infants wrapped their fingers firmly around the ends of the dumbbell and gave it a hard yank, successfully pulling it apart. Further work revealed the social nature of the process. An inanimate device was built, and the infants watched as mechanical pincers slipped off the ends of the same object in the same way the human fingers had done. When given the object, the infants picked it up but did *not* try to reenact the target act. An inference from this and related experiments is that infants have a primitive "theory of mind" that gives them the capacity to reason about the adult's goals rather than just their surface behavior.[9] Children can pick up the nonobvious or hidden messages that lie behind adult actions.

Building on these studies about the power of mere observation of adult behavior to spark children's actions, we now turn to work directly addressing children's acquisition of social biases. Modern research in child development has shown that young children exhibit biases based on race, gender, language, and other attributes during the preschool and elementary school years.[10] We therefore wanted to look at preschool children to understand in more detail how novel biases might first be created in the mind of the young child.[11]

One study presented four- to five-year-old preschoolers with video clips of adult biased behaviors.[12] The videos showed interactions between one adult (the "actor") and two other adults (the "targets"). The targets sat on either side of the central actor and wore differently colored T-shirts to distinguish one from the other. The video scenario showed the actor handing each of the targets a toy. While distributing the toy, the actor exuded positive nonverbal signals toward one of the targets (smiling, leaning in, using a warm tone of voice) and negative nonverbal signals toward the other target (scowling, leaning away, using a cold tone of voice) – conveying bias through action. The thirty-second script was played twice, each time with a different central actor. The same target received the negative (or positive) signals from both actors, indicating that more than one person

held the negative (or positive) attitude toward these targeted people. Preschoolers were transfixed by this video of adult interaction, looking back and forth between the actor and the targets as the script unfolded.

We then assessed children's attitudes, cognition, and behavior toward the two targets, using social preferences (who they liked), their distribution of resources (sharing an attractive toy), and who they imitated (who they were willing to learn from). The results showed that preschoolers treated the targets differentially. Children adopted attitudes and behaviors that strongly favored the target of positive nonverbal signals relative to the target of negative nonverbal signals.

The key question arising from the results: can witnessing biased adult behavior directed toward a person of a certain color (here a particular T-shirt color) create social bias in the child that generalizes to a group of others who "look like" this negatively targeted individual? In a new study, children's responses were assessed using pictures of two groups of novel people.[13] One group of new people wore the same-colored T-shirt as the target of the negative signal; the other group of new people wore the same-colored T-shirt as the target of the positive signal. The social preference and imitation tasks were repeated, and a new measure was included to assess which social group the child would choose to play with when they were told that another person had to be added to the game. Preschoolers displayed a bias toward liking, imitating, and wanting to play with social group members who looked like the targets of positive signals. They passed on the chance to play with someone from the negatively marked group, which – if it occurred in real life – would translate into reduced interactions with those from the disfavored group.

Let's now turn our attention to examples of opportunities for observational social learning in the everyday lives of young children, specifically within their early care and education programs. In 2019, almost 60 percent of young children in the United States were enrolled in some form of nonfamilial care arrangement, often beginning in infancy or toddlerhood.[14] While attending these early care and education programs, young children see and experience a myriad of adult social interactions, including how adult caregivers interact with other young children in the classroom and other adult caregivers, such as assistant teachers. We argue that these adult social interactions provide opportunities for young children to observe and learn implicit racial biases.

Expulsion and suspension from early care and education programs present a powerful, yet often unintentional, opportunity for young children to observe racial disparities in preschool. In the first national study of preschool expulsion rates, conducted between 2002 and 2004, preschoolers (three to four years old) were found to be expelled at a rate of more than three times that of K–12 students, and the rates in community-based child care programs were even higher.[15] Black preschoolers were more than twice as likely to be expelled as white preschoolers,

and boys were more than four times as likely to be expelled as girls, despite no evidence that either Black children or boys exhibited greater levels of misbehavior. Race and gender were found to interact, yielding especially high rates of expulsion for Black boys. In mixed-age classrooms, older preschoolers were more likely to be expelled than their younger classmates, which we speculatively link to other research demonstrating a tendency to view Black children as being older and more threatening.[16] Similar race and gender disparities were found in more recent studies conducted by the U.S. Department of Education's Office of Civil Rights (OCR).[17] In June 2016, the OCR reported that Black preschoolers were 3.6 times as likely to be suspended as white preschoolers. Although Black preschoolers represented only 19 percent of the preschool population, they received 47 percent of suspensions; and although boys represented 54 percent of all preschoolers, they received 78 percent of all suspensions. Fortunately, twenty-nine states, plus the federal Head Start program, now have legislation or executive branch policies aimed at limiting expulsions and suspensions in early care and education settings, with virtually all policy actions initiated since 2015.[18]

Nonetheless, early expulsions and suspensions predict a host of negative life outcomes, likely because of a resulting damaged relationship to schools/teachers and a concomitant denial of educational opportunities. The potential pathways to subsequent negative life outcomes include negative school attitudes, academic failure and grade retention, later suspensions and expulsions, a tenfold increased likelihood of high school dropout, and an eightfold increased likelihood of adult incarceration.[19] Further suggesting potential correlations between early expulsions and later incarceration, the rate at which U.S. early care and education programs expel young children (6.7 to 7.5 per 1,000) is similar to the rate at which adults are incarcerated (6.4 per 1,000), with similar levels of race and gender disparities.[20] Preschool-age children of incarcerated adults are at a threefold increased likelihood of themselves being expelled from preschool, painting a picture of racially disproportional intergenerational exclusion.[21]

The poignant irony of the disproportional expulsion and suspension of Black preschoolers is that the initial argument for early care and education programs in the United States used research data obtained from overwhelmingly Black preschoolers and their families. The three most widely cited studies used to build the case for early care and education (Perry Preschool Study, Abecedarian Study, and Chicago Child-Parent Centers Study) were conducted with child participant samples that were, respectively, 100 percent, 98 percent, and 93 percent Black.[22] Nonetheless, it is Black preschoolers who are most likely to be excluded from these same programs through racially disproportional expulsion and suspension practices.[23]

Racial disproportionality in early childhood expulsions and suspensions has serious downstream implications. Why does it exist? Early work in K–12 schools focused on adult biases about children of color and their behaviors, particularly

Black boys. Using elementary school disciplinary records, psychologist Russell J. Skiba and colleagues found that Black boys were more likely to be suspended or expelled relative to other students, even when the behaviors cited as the reason for the disciplinary sanctions were similar in severity.[24] Relatedly, other studies have documented adult biases that viewed Black boys as being more likely to engage in misbehavior, as well as a tendency for adults to overestimate the age of Black boys and view them as bigger and more dangerous.[25] Although such biases regarding Black boys may contribute to the extreme racial disparities in suspension and expulsion rates, the work is correlational and few studies have been designed to directly measure race/gender biases that might be exhibited by early educators toward preschoolers during everyday activities.

In the first such study, a high-tech eye-tracking device was employed to assess whether preschool teachers might assume and anticipate a greater likelihood of disruptive behaviors from Black preschoolers (especially Black boys) relative to white preschoolers.[26] The participants included one hundred seventeen early educators from around the United States attending a national early childhood education conference. Teachers were seated in front of a fifteen-inch laptop computer screen equipped with an eye-tracking device that was calibrated to their gaze and capable of measuring where they were looking on the screen. The participants were shown twelve thirty-second videos of four preschoolers (each four years old) in an early education classroom: a Black boy, a Black girl, a white boy, and a white girl. The twelve videos were recorded from different angles to balance the location of each child on the screen, one angle is shown on the left side of Figure 1. The participants were instructed to watch the videos, look for evidence of "challenging behaviors," and press a keypad button whenever they saw a behavior that could turn into a "challenging behavior" – all while their eye gaze was precisely tracked so that the amount of time the teachers spent looking at each child could be tabulated and analyzed.

At the end of the videos, the teachers were shown a picture of each of the four preschoolers, as displayed on the right side of Figure 1, and were asked which one of the preschoolers they felt required the most of their attention. The participants were not told until after the study that there was actually *no* challenging behavior in the videos: all four preschoolers were child actors who simply played with the objects as directed.

Results indicated that early educators, when expecting challenging behaviors, spent significantly more time focusing their gaze on the Black preschoolers, especially the Black boy. This finding was consistent regardless of the race of the teacher. However, when teachers were asked explicitly which child they believed required the most of their attention, results indicated that teachers believed they were most closely watching for misbehavior based on gender. The most common

Figure 1
Illustrations of the Video from the Eye-Tracking Study Done with Early Care and Education Teachers

The image on the left shows four children (two white, two Black, two girls, two boys) playing with objects. The children were child actors who showed no misbehavior during the video. The image on the right is the final screen presented to the teachers, who were asked for an explicit self-report of which child required the most attention. Source: Walter S. Gilliam, Angela N. Maupin, Chin R. Reyes, et al., "Do Early Educator's Implicit Biases Regarding Sex and Race Relate to Behavior Expectations and Recommendations of Preschool Expulsions and Suspensions?" paper presented at the U.S. Administration for Children and Families State and Territory Administrators Meeting, Alexandria, Va., September 28, 2016, https://marylandfamiliesengage.org/wp-content/uploads/2019/07/Preschool-Implicit-Bias-Policy-Brief.pdf.

response was the Black boy (42 percent), followed by the white boy (34 percent), white girl (13 percent), and Black girl (10 percent). Either way, the Black boy ended up with the short end of the stick.[27]

In short, when early educators were led to believe that a preschooler might misbehave, they focused their attention more acutely on the Black boy, anticipating bad behavior that never was to happen. In a way, this study resembles the first day of preschool for early educators. Teachers are presented with a group of preschoolers they have never met and are placed in a position where they might make assumptions about what kinds of behaviors to expect from each of them. If the eye-tracking study reflects to some degree how early educators behave in a real classroom, race and gender biases could account for at least some of the disproportional rates of preschool suspensions and expulsions of Black boys.[28]

But how do the *other* preschoolers in the classroom experience this extra vigilance placed on the Black boys in the classroom? As discussed earlier, young children are astute observers of adult behavior, and even subtle displays of negative affect and attention by adults are noted by young children who then shape their own biases based on this observational social learning. While the adults are focusing most acutely on the Black boys when expecting misbehavior, the other chil-

dren in the room are watching the adults closely and actively forming their own expectations and biases based on those observations.[29] Sometimes the lessons that stick the most are the ones never intended to be taught.

Preschool children are also being provided, unintentionally, with hundreds of hours of "data" about implicit racial bias at an *institutional* level, beyond the acts of individual teachers. Most early care and education programs have more than one adult in the room at the same time, often a lead teacher plus one or more assistant teachers or aides. These assistant teachers and aides are more likely than lead teachers to identify as a person of color (41 percent versus 35 percent in center-based programs and 66 percent versus 30 percent in home-based programs).[30]

The roles that the assistant teachers and aides play and the duties they perform are supportive, but usually quite different than those performed by lead teachers. In a nationally representative study of 3,191 preschool classrooms, assistant teachers were reported as being more likely to perform duties such as cleaning tables and setting up rest areas in the room, rather than leading the teaching of the children, working with parents, or providing overall planning activities.[31] This was especially true when there was a relatively larger discrepancy in educational level between the lead and assistant teacher. Taken together, white early care and education staff are more likely to engage in higher-paid/higher-status tasks, while nonwhite staff are performing lower-status tasks under the lead teachers' supervision, communicating ideas about who has power and authority – and children watch this daily.

Additionally, an emerging body of evidence suggests that a race/ethnicity match between students and teachers may be beneficial to young children of color. In an eleven-state study of more than seven hundred prekindergarten classrooms, Hispanic/Latinx preschoolers scored higher on academic skills when in classrooms where the teacher was also Hispanic/Latinx, and Black preschoolers scored higher on teacher-reported academic and social-emotional measures when the teacher was Black.[32] These findings, although only correlational, are similar to those found in a study of elementary school students.[33]

Although such studies have led to increased calls to diversify further the early childhood teacher workforce, care must be taken to ensure that early educators of color are seen by the young children in their classrooms as taking an active and vital role in their education and care.[34] Otherwise, young children may be exposed, unintentionally, to racially defined social hierarchies within the early education setting, in which children are more likely to see white adults in leadership roles and adults of color in more subordinate roles. Child development research shows that young children are finely attuned to cues about prestige, power, and social status.[35] Although racialized patterns of job responsibilities and leadership op-

portunities are common across many employment settings,[36] when this happens in early care and education programs it provides another pathway by which implicit biases may begin to inform children's racialized expectations about social roles.[37]

Finally, early educators themselves may be the targets of racial bias, and their experiences of racism can impact the quality of care they provide. Recent findings show that during the height of the ongoing COVID-19 pandemic, early care and education professionals of color have experienced high rates of racialized aggression in their own daily lives, which is associated with increased experiences of stress particularly in Black and Asian early educators.[38] Both job stress and depression in early educators have been associated with increased rates of early childhood expulsions and suspensions, providing yet another pathway by which racial biases may increase the rate of early childhood exclusions (which have consistently been shown to be applied in racially disproportional rates).[39]

What practical steps can be taken based on scientific studies of young children and their experiences in bias-rich, real-world settings? Much has been written about the mixed results of attempts to remedy implicit racial biases in adults.[40] One wonders whether more positive results might be obtained through intervention programs designed for young children.[41] In other domains of child development, early identification and treatment are more effective and less costly than interventions at older ages when a habit or attribute is more entrenched.[42] This may be overly optimistic in the case of racial bias, because children will inevitably be exposed to pervasive racial inequities as they grow up, which could overwhelm a short-lived treatment program. Yet it is not inconceivable that efforts could alter some of these environmental conditions for children. In our experience, parents and early educators alike are taken aback to hear that young children's mere exposure to adult nonverbal behavior patterns (which often contain unintentionally biased behaviors) can subsequently influence the children's own behavior toward others.[43] It may be useful to move these and future scientific findings more rapidly into the hands of parents and early educators.

What might a parent do with this information? It is not impossible that parents could regulate their own behavior while in the presence of their own children. For example, there are white-Black differences in the frequency and content of parents' conversations about race with their children.[44] White parents are often uncertain about whether to engage with conversations about race, concerned about what messages are age appropriate, and anxious that such conversations may focus their child on racial differences that inadvertently stoke racial prejudice and implicit racial biases.[45] Although some white parents may attempt to become "colorblind" and "color mute" in interactions with their children, this may im-

plicitly convey that race is a taboo topic, leaving young children to fill in the gaps from other sources (media, rumors from peers, chance encounters). The most feasible, wise, and efficacious recommendations for parents about discussing race with children are not well understood, and the topic deserves more research.[46]

Other information of potential interest to parents is that fostering intergroup contact and friendships with children from another race is a promising avenue for reducing racial bias in children.[47] However, in contemporary U.S. society, neighborhoods, K–12 schools, and early care and education programs exhibit a high degree of segregation. More than one-third of all K–12 students attend a school in which 75 percent or more of the students are of a single racial group, and early care and education programs are considerably *more* racially segregated.[48] Schools in the United States are more racially segregated than their surrounding neighborhoods, and so too are early childhood programs.[49] Even parents who explicitly avow egalitarian views do not necessarily bring other-race acquaintances into their homes on a regular basis. Parents might make different choices about their own behavior and their children's playdates if given information about potential benefits of intergroup contact and friendships for their own child. Further applied-science studies in this area would be valuable.

It is also time to think about bringing more scientific information to colleges of education and teacher internships. It may be eye-opening for early care and education teachers to learn that their own implicit biases leak out in the classroom. Young children are acutely attentive to the direction of adult gaze, and teachers look toward Black boys when they anticipate trouble. It remains unknown whether information about implicit bias in the classroom could be used to enhance educational equity if effectively conveyed to teachers.[50] The National Academy of Education is attuned to these issues and is seeking to incorporate lessons on civics in an expanded U.S. educational agenda.[51] There is also a push by the National Academy of Education and other psychology-focused organizations to translate scientific research into practical actions to improve the educational experience and foster the opportunity to thrive for all children. Among other things, this convergence is spotlighting the urgent need for increased understanding of the mechanisms by which racial biases are transmitted to children, often unintentionally, both in and out of school – and what might be done about it.

In conclusion, young children are social pattern detectors. They study our behavior, and sometimes the nonverbal messages they receive are not the ones we intend to send. What every parent, teacher, and societal leader should think about is that children watch and learn from our behavior before first grade. When we exhibit biases in front of young children, we are unwittingly instilling our biases in their minds – biases they then adopt, practice, and perpetuate.

AUTHORS' NOTE

The work on this essay was supported by funds from the Bezos Family Foundation. We thank Allison Skinner-Dorkenoo, Leoandra Onnie Rogers, Sapna Cheryan, Patricia Kuhl, and Goodwin Liu for thoughtful comments on an earlier draft.

ABOUT THE AUTHORS

Andrew N. Meltzoff, a Fellow of the American Academy since 2009, is the Co-Director of the Institute for Learning & Brain Sciences and Professor of Psychology at the University of Washington, where he holds the Job and Gertrud Tamaki Endowed Chair. He is the author of *Words, Thoughts, and Theories* (with Alison Gopnik, 1997) and *The Scientist in the Crib: What Early Learning Tells Us About the Mind* (with Alison Gopnik and Patricia Kuhl, 1999), and the editor of *The Imitative Mind: Development, Evolution, and Brain Bases* (with Wolfgang Prinz, 2002).

Walter S. Gilliam is Executive Director of the Buffett Early Childhood Institute at the University of Nebraska, where he holds the Richard D. Holland Presidential Chair in Early Childhood Development, and Professor in the Munroe-Meyer Institute at the University of Nebraska Medical Center. He is also Professor Adjunct in the Child Study Center at the Yale School of Medicine, a Senior Fellow at the Bipartisan Policy Center, President of Zero to Three, and a former President of Child Care Aware of America. He is the author of *A Vision for Universal Preschool Education* (with Edward Zigler and Stephanie Jones, 2006).

ENDNOTES

[1] Bruno Lasker, *Race Attitudes in Children* (New York: Henry Holt, 1929); and Gordon W. Allport, *The Nature of Prejudice* (Reading, Mass.: Addison-Wesley, 1954).

[2] Kenneth B. Clark and Mamie P. Clark, "Racial Identification and Preference in Negro Children," in *Readings in Social Psychology*, ed. Theodore M. Newcomb and Eugene L. Hartley (New York: Henry Holt, 1947), 602–611.

[3] Rebecca A. Dore, Kelly M. Hoffman, Angeline S. Lillard, and Sophie Trawalter, "Children's Racial Bias in Perceptions of Others' Pain," *British Journal of Developmental Psychology* 32 (2) (2014): 218–231, https://doi.org/10.1111/bjdp.12038.

[4] Patricia K. Kuhl, "Early Language Acquisition: Cracking the Speech Code," *Nature Reviews Neuroscience* 5 (11) (2004): 831–843, https://doi.org/10.1038/Nrn1533; and Janet F. Werker and Takao K. Hensch, "Critical Periods in Speech Perception: New Directions," *Annual Review of Psychology* 66 (1) (2015): 173–196, https://doi.org/10.1146/annurev-psych-010814-015104.

[5] Alison Gopnik and Andrew N. Meltzoff, *Words, Thoughts, and Theories* (Cambridge, Mass.: The MIT Press, 1997); Susan A. Gelman, *The Essential Child: Origins of Essentialism in Everyday Thought* (New York: Oxford University Press, 2003); Michael Tomasello, *Becoming Human: A Theory of Ontogeny* (Cambridge, Mass.: Harvard University Press, 2019); and Elizabeth Spelke, *What Babies Know* (New York: Oxford University Press, 2022).

[6] Andrew N. Meltzoff, Patricia K. Kuhl, Javier Movellan, and Terrence J. Sejnowski, "Foundations for a New Science of Learning," *Science* 325 (5938) (2009): 284–288, https://doi.org/10.1126/science.1175626; Andrew N. Meltzoff and Peter J. Marshall, "Human Infant Imitation as a Social Survival Circuit," *Current Opinion in Behavioral Sciences* 24 (2018): 130–136, https://doi.org/10.1016/j.cobeha.2018.09.006; Tomasello, *Becoming Human*; and Andrew N. Meltzoff and Peter J. Marshall, "Importance of Body Representations in Social-Cognitive Development: New Insights from Infant Brain Science," *Progress in Brain Research* 254 (2020): 25–48, https://doi.org/10.1016/bs.pbr.2020.07.009.

[7] Pamela J. Klein and Andrew N. Meltzoff, "Long-Term Memory, Forgetting, and Deferred Imitation in 12-Month-Old Infants," *Developmental Science* 2 (1) (1999): 102–113, https://doi.org/10.1111/1467-7687.00060.

[8] Andrew N. Meltzoff, "Understanding the Intentions of Others: Re-enactment of Intended Acts by 18-Month-Old Children," *Developmental Psychology* 31 (5) (1995): 838–850, https://doi.org/10.1037/0012-1649.31.5.838.

[9] Amanda L. Woodward, "Infants' Grasp of Others' Intentions," *Current Directions in Psychological Science* 18 (1) (2009): 53–57, https://doi.org/10.1111/j.1467-8721.2009.01605.x; Andrew N. Meltzoff, "Social Cognition and the Origins of Imitation, Empathy, and Theory of Mind," in *The Wiley-Blackwell Handbook of Childhood Cognitive Development*, ed. Usha Goswami (Malden, Mass.: Wiley-Blackwell, 2011), 49–75; and Andrew N. Meltzoff, Anna Waismeyer, and Alison Gopnik, "Learning About Causes From People: Observational Causal Learning in 24-Month-Old Infants," *Developmental Psychology* 48 (5) (2012): 1215–1228, https://doi.org/10.1037/a0027440.

[10] Andrew S. Baron and Mahzarin R. Banaji, "The Development of Implicit Attitudes: Evidence of Race Evaluations from Ages 6 and 10 and Adulthood," *Psychological Science* 17 (1) (2006): 53–58, https://doi.org/10.1111/j.1467-9280.2005.01664.x; Allison L. Skinner and Andrew N. Meltzoff, "Childhood Experiences and Intergroup Biases Among Children," *Social Issues and Policy Review* 13 (1) (2019): 211–240, https://doi.org/10.1111/sipr.12054; Steven O. Roberts and Michael T. Rizzo, "The Psychology of American Racism," *American Psychologist* 76 (3) (2021): 475–487, http://doi.org/10.1037/amp0000642; Sandra R. Waxman, "Racial Awareness and Bias Begin Early: Developmental Entry Points, Challenges, and a Call to Action," *Perspectives on Psychological Science* 16 (5) (2021): 893–902, https://doi.org/10.1177/17456916211026968; Allison L. Skinner-Dorkenoo, Meghan George, James E. Wages III, et al., "A Systemic Approach to the Psychology of Racial Bias Within Individuals and Society," *Nature Reviews Psychology* 2 (2023): 392–406, https://doi.org/10.1038/s44159-023-00190-z; and Diane Hughes, Blair Cox, and Sohini Das, "Growing Up, Learning Race: An Integration of Research on Cognitive Mechanisms and Socialization in Context," *Annual Review of Developmental Psychology* 5 (2023): 137–167, https://doi.org/10.1146/annurev-devpsych-120321-015718.

[11] Three-month-old infants prefer to look at faces from their own racial group, but this preference does not exist in newborns, suggesting that experience plays a role in forming such preferences. See Gizelle Anzures, Paul C. Quinn, Olivier Pascalis, et al., "Developmental Origins of the Other-Race Effect," *Current Directions in Psychological Science* 22 (3) (2013): 173–178, https://doi.org/10.1177/0963721412474459. The racial biases and prejudices of preschoolers and older children implicate something above and beyond visual preferences (or visual categorization). They involve differential beliefs, attitudes, and behaviors directed toward people. More work is needed to map connections between these developmental levels. See Kang Lee, Paul C. Quinn, and Olivier Pascalis, "Face Race Processing and Racial Bias in Early Development: A Perceptual-

Social Linkage," *Current Directions in Psychological Science* 26 (3) (2017): 256–262, https://doi.org/10.1177/0963721417690276; and Waxman, "Racial Awareness and Bias Begin Early."

[12] Allison L. Skinner, Andrew N. Meltzoff, and Kristina R. Olson, "'Catching' Social Bias: Exposure to Biased Nonverbal Signals Creates Social Biases in Preschool Children," *Psychological Science* 28 (2) (2017): 216–224, https://doi.org/10.1177/0956797616678930.

[13] Allison L. Skinner, Kristina R. Olson, and Andrew N. Meltzoff, "Acquiring Group Bias: Observing Other People's Nonverbal Signals Can Create Social Group Biases," *Journal of Personality and Social Psychology* 119 (4) (2020): 824–838, https://doi.org/10.1037/pspi0000218.

[14] See "Table 1. Percentage of Children from Birth through Age 5 and Not Yet in Kindergarten Participating in Various Weekly Nonparental Care Arrangements, by Child and Family Characteristics: 2019" in Jiashan Cui, Luke Natzke, and Sarah Grady, *Early Childhood Program Participation: 2019*, NCES 2020-075REV (Washington, D.C.: United States Department of Education, 2021), A-1–A-3, https://nces.ed.gov/pubs2020/2020075REV.pdf.

[15] Walter S. Gilliam, *Prekindergarteners Left Behind: Expulsion Rates in State Prekindergarten Systems* (New Haven, Conn.: Yale University, 2005); and Walter S. Gilliam and Golan Shahar, "Preschool and Child Care Expulsion and Suspension: Rates and Predictors in One State," *Infants and Young Children* 19 (3) (2006): 228–245, https://doi.org/10.1097/00001163-200607000-00007.

[16] Phillip A. Goff, Matthew Christian Jackson, Brooke Allison Lewis Di Leone, et al., "The Essence of Innocence: Consequences of Dehumanizing Black Children," *Journal of Personality and Social Psychology* 106 (4) (2014): 526–545, https://doi.org/10.1037/a0035663.

[17] U.S. Department of Education Office of Civil Rights, "Civil Rights Data Collection: Data Snapshot: Early Childhood Education" (Washington, D.C.: U.S. Department of Education, 2014), https://www2.ed.gov/about/offices/list/ocr/docs/crdc-early-learning-snapshot.pdf; and U.S. Department of Education Office of Civil Rights, "2013–2014 Civil Rights Data Collection: Key Data Highlights on Equity and Opportunity Gaps in Our Nation's Public Schools" (Washington, D.C.: U.S. Department of Education, 2016), http://www2.ed.gov/about/offices/list/ocr/docs/crdc-2013-14.html.

[18] Alysse Melville Loomis, Annie Davis, Gracelyn Cruden, et al., "Early Childhood Suspension and Expulsion: A Content Analysis of State Legislation," *Early Childhood Education Journal* 50 (2) (2022): 327–344, https://doi.org/10.1007/s10643-021-01159-4; and Carey McCann, Sheila Smith, Uyen (Sophie) Nguyen, and Maribel R. Granja, *States' Growing Commitment to Preventing Young Children's Expulsion from Early Care and Education Programs: Results of a 50-State Policy Survey* (New York: National Center for Children in Poverty, 2021), https://www.nccp.org/wp-content/uploads/2021/10/States-Growing-Commitment-to-Preventing-Young-Childrens-Expulsion-from-ECE-Programs.pdf.

[19] American Psychological Association Zero Tolerance Task Force, "Are Zero Tolerance Policies Effective in the Schools? An Evidentiary Review and Recommendations," *American Psychologist* 63 (9) (2008): 852–862, https://doi.org/10.1037/0003-066X.63.9.852; Hanno Petras, Katherine E. Masyn, Jacquelyn A. Buckley, et al., "Who is Most at Risk for School Removal? A Multilevel Discrete-Time Survival Analysis of Individual- and Context-Level Influences," *Journal of Educational Psychology* 103 (1) (2011): 223–237, https://doi.org/10.1037/a0021545; and Council on School Health, "Out-of-School Suspension

and Expulsion," *Pediatrics* 131 (3) (2013): e1000–e1007, https://doi.org/10.1542/peds.2012-3932.

[20] Gilliam, *Prekindergarteners Left Behind*; Jacob Kang-Brown, Chase Montagnet, and Jasmine Heiss, *People in Jail and Prison in 2020* (New York: Vera Institute of Justice, 2021), https://www.vera.org/downloads/publications/people-in-jail-and-prison-in-2020.pdf; Songtian Zeng, Catherine P. Corr, Courtney O'Grady, and Yiyang Guan, "Adverse Childhood Experiences and Preschool Suspension Expulsion: A Population Study," *Child Abuse & Neglect* 97 (2019): 104149, https://doi.org/10.1016/j.chiabu.2019.104149; U.S. Department of Education Office of Civil Rights, "2013–2014 Civil Rights Data Collection"; and Zhen Zeng and Todd D. Minton, *Jail Inmates in 2019* (Washington, D.C.: U.S. Department of Justice, Bureau of Justice Statistics, 2021), https://bjs.ojp.gov/redirect-legacy/content/pub/pdf/ji19.pdf.

[21] Zeng, Corr, O'Grady, and Guan, "Adverse Childhood Experiences and Preschool Suspension Expulsion."

[22] Lawrence J. Schweinhart, Jeanne Montie, Zongping Xiang, et al., *Lifetime Effects: The High/Scope Perry Preschool Study through Age 40* (Ypsilanti, Mich.: High/Scope Press, 2005), 194–215, https://nieer.org/wp-content/uploads/2014/09/specialsummary_rev2011_02_2.pdf; Frances A. Campbell and Craig T. Ramey, "Carolina Abecedarian Project," in *Childhood Programs and Practices in the First Decade of Life: A Human Capital Integration*, ed. Arthur J. Reynolds, Arthur J. Rolnick, Michelle M. Englund, and Judy A. Temple (New York: Cambridge University Press, 2010), 76–98, https://doi.org/10.1017/CBO9780511762666.005; and Arthur J. Reynolds, Suh-Ruu Ou, and Judy A. Temple, "A Multicomponent, Preschool to Third Grade Preventive Intervention and Educational Attainment at 35 Years of Age," *JAMA Pediatrics* 172 (3) (2018): 247–256, https://doi.org/10.1001/jamapediatrics.2017.4673.

[23] This point was first made in congressional testimony to the House of Representatives Committee on Appropriations, Subcommittee on Labor, Health and Human Services, Education, and Related Services, Budget Hearing–Early Education Panel, by Walter S. Gilliam on April 14, 2015. See House Appropriations Committee, "Hearing: Early Education Panel," YouTube video streamed live on April 14, 2015, https://www.youtube.com/watch?v=Dd39yX-P5VA.

[24] Russell J. Skiba, Robert H. Horner, Choong-Geun Chung, et al., "Race Is Not Neutral: A National Investigation of African American and Latino Disproportionality in School Discipline," *School Psychology Review* 40 (1) (2011): 85–107, https://doi.org/10.1080/02796015.2011.12087730.

[25] Jennifer L. Eberhardt, Phillip Atiba Goff, Valerie J. Purdie, and Paul G. Davies, "Seeing Black: Race, Crime, and Visual Processing," *Journal of Personality and Social Psychology* 87 (6) (2004): 876–893, https://doi.org/10.1037/0022-3514.87.6.876; Goff, Jackson, Di Leone, et al., "The Essence of Innocence"; Jennifer L. Eberhardt and Jason A. Okonofua, "Two Strikes: Race and the Disciplining of Young Students," *Psychological Science* 26 (5) (2015): 617–624, https://doi.org/10.1177/0956797615570365; Andrew R. Todd, Kelsey C. Thiem, and Rebecca Neel, "Does Seeing Faces of Young Black Boys Facilitate the Identification of Threatening Stimuli?" *Psychological Science* 27 (3) (2016): 384–393, https://doi.org/10.1177/0956797615624492; and Andrew R. Todd, Austin J. Simpson, Kelsey C. Thiem, and Rebecca Neel, "The Generalization of Implicit Racial Bias to Young Black Boys: Automatic Stereotyping or Automatic Prejudice?" *Social Cognition* 34 (4) (2016): 306–323, https://doi.org/10.1521/soco.2016.34.4.306.

[26] Walter S. Gilliam, Angela N. Maupin, Chin R. Reyes, et al., "Do Early Educator's Implicit Biases Regarding Sex and Race Relate to Behavior Expectations and Recommendations of Preschool Expulsions and Suspensions?" paper presented at the U.S. Administration for Children and Families State and Territory Administrators Meeting, Alexandria, Va., September 28, 2016, https://marylandfamiliesengage.org/wp-content/uploads/2019/07/Preschool-Implicit-Bias-Policy-Brief.pdf.

[27] Bias against Black boys in particular has also been shown in laboratory experiments with preschoolers using both implicit and explicit measures. See Danielle R. Perszyk, Ryan F. Lei, Galen V. Bodenhausen, et al., "Bias at the Intersection of Race and Gender: Evidence From Preschool-Aged Children," *Developmental Science* 22 (3) (2019): e12788, https://doi.org/10.1111/desc.12788.

[28] Terri J. Sabol, Courtenay L. Kessler, Leoandra Onnie Rogers, et al., "A Window into Racial and Socioeconomic Status Disparities in Preschool Disciplinary Action Using Developmental Methodology," *Annals of the New York Academy of Sciences* 1508 (1) (2022): 123–136, https://doi.org/10.1111/nyas.14687.

[29] Elizabeth Brey and Kristin Shutts, "Children Use Nonverbal Cues From an Adult to Evaluate Peers," *Journal of Cognition and Development* 19 (2) (2018): 121–136, https://doi.org/10.1080/15248372.2018.1449749; and Allison Master, Andrew N. Meltzoff, and Sapna Cheryan, "Gender Stereotypes about Interests Start Early and Cause Gender Disparities in Computer Science and Engineering," *Proceedings of the National Academy of Sciences* 118 (48) (2021): e2100030118, https://doi.org/10.1073/pnas.2100030118.

[30] Wayne Mayfield and Ikhee Cho, *The National Workforce Registry Alliance 2021 Workforce Dataset: Early Childhood and School-Age Workforce Trends, with a Focus on Racial/Ethnic Equity* (Washington, D.C.: National Workforce Registry Alliance, 2022), https://www.registryalliance.org/wp-content/uploads/2022/05/NWRA-2022-ECE-workforce-data-report-final.pdf; and Katherine Paschall, Rebecca Madill, and Tamara Halle, *Demographic Characteristics of the Early Care and Education Workforce: Comparisons with Child and Community Characteristics* (Washington, D.C.: Office of Planning, Research, and Evaluation, Administration for Children and Families, United States Department of Health and Human Services, 2020), https://www.acf.hhs.gov/sites/default/files/documents/opre/demographic-characteristics-ECE-dec-2020.pdf.

[31] Laura S. Sosinsky and Walter S. Gilliam, "Assistant Teachers in Prekindergarten Programs: What Roles Do Lead Teachers Feel Assistants Play in Classroom Management and Teaching?" *Early Education and Development* 22 (4) (2011): 676–706, https://doi.org/10.1080/10409289.2010.497432.

[32] Jason T. Downer, Priscilla Goble, Sonya S. Myers, and Robert C. Pianta, "Teacher-Child Racial/Ethnic Match within Pre-Kindergarten Classrooms and Children's Early School Adjustment," *Early Childhood Research Quarterly* 37 (4) (2016): 26–38, https://doi.org/10.1016/j.ecresq.2016.02.007.

[33] Damira S. Rasheed, Joshua L. Brown, Sebrina L. Doyle, and Patricia A. Jennings, "The Effect of Teacher–Child Race/Ethnicity Matching and Classroom Diversity on Children's Socioemotional and Academic Skills," *Child Development* 91 (3) (2020): e597–e618, https://doi.org/10.1111/cdev.13275.

[34] Sandra L. Soliday Hong, Kamilah B. Legette, Laura Kuhn, et al., "Lead Teacher, Assistant Teacher, and Peer Racial/Ethnic Match and Child Outcomes for Black Children En-

rolled in Enhanced High-Quality Early Care and Education Programs," *Early Childhood Research Quarterly* 64 (3) (2023): 186–198, https://doi.org/10.1016/j.ecresq.2023.03.001.

[35] Anna-Kaisa Newheiser, Yarrow Dunham, Anna Merrill, et al., "Preference for High Status Predicts Implicit Outgroup Bias Among Children from Low-Status Groups," *Developmental Psychology* 50 (4) (2014): 1081–1090, https://doi.org/10.1037/a0035054; Anthea Pun, Susan A. J. Birch, and Andrew S. Baron, "Foundations of Reasoning about Social Dominance," *Child Development Perspectives* 11 (3) (2017): 155–160, http://doi.org/10.1111/cdep.12235; Selin Gülgöz and Susan Gelman, "Who's the Boss: Concepts of Social Power across Development," *Child Development* 88 (3) (2017): 946–963, https://doi.org/10.1111/cdev.12643; Ashley J. Thomas, Lotte Thomsen, Angela F. Lukowski, et al., "Toddlers Prefer Those Who Win But Not When They Win By Force," *Nature Human Behaviour* 2 (9) (2018): 662–669, https://doi.org/10.1038/s41562-018-0415-3; and Isobel A. Heck, Kristin Shutts, and Katherine D. Kinzler, "Children's Thinking about Group-Based Social Hierarchies," *Trends in Cognitive Sciences* 26 (7) (2022): 593–606, https://doi.org/10.1016/j.tics.2022.04.004.

[36] Mladen Adamovic and Andreas Leibbrandt, "Is There a Glass Ceiling for Ethnic Minorities to Enter Leadership Positions? Evidence from a Field Experiment with over 12,000 Job Applications," *The Leadership Quarterly* 34 (2) (2023): 101655, https://doi.org/10.1016/j.leaqua.2022.101655.

[37] We have highlighted examples of institutional and systemic bias that impact children in the preschool environment. We acknowledge that additional sources of systemic bias are "in the air," which influence very young children and K–12 students, often without awareness or deliberate intent. See Roberts and Rizzo, "The Psychology of American Racism"; Tessa E. S. Charlesworth, Victor Yang, Thomas C. Mann, et al., "Gender Stereotypes in Natural Language: Word Embeddings Show Robust Consistency Across Child and Adult Language Corporal of More Than 65 Million Words," *Psychological Science* 32 (2) (2021): 218–240, https://doi.org/10.1177/0956797620963619; Skinner-Dorkenoo, George, Wages III, et al., "A Systemic Approach to the Psychology of Racial Bias"; Hughes, Cox, and Das, "Growing Up, Learning Race"; Manuel J. Galvan and B. Keith Payne, "Implicit Bias as a Cognitive Manifestation of Systemic Racism," *Dædalus* 153 (1) (Winter 2024): 106–122, https://www.amacad.org/publication/implicit-bias-cognitive-manifestation-systemic-racism; and Rebecca C. Hetey, MarYam G. Hamedani, Hazel Rose Markus, and Jennifer L. Eberhardt, "'When the Cruiser Lights Come On': Using the Science of Bias & Culture to Combat Racial Disparities in Policing," *Dædalus* 153 (1) (Winter 2024): 123–150, https://www.amacad.org/publication/when-cruiser-lights-come-using-science-bias-culture-combat-racial-disparities-policing.

[38] Ayse Cobanoglu and Walter S. Gilliam, "Double Pandemic: Exposure to Racial Aggression and Well-Being of Early Child Care Providers," in Sara Vecchiotti (Chair), *Understanding Workforce Wellbeing during the COVID-19 Pandemic: Confronting Trauma and Racism in Early Care and Education*, paper presented at the National Research Conference on Early Childhood, U.S. Administration for Children and Families, Washington, D.C., July 29, 2022.

[39] Gilliam and Shahar, "Preschool and Child Care Expulsion and Suspension: Rates and Predictors in One State."

[40] Elizabeth Levy Paluck, Roni Porat, Chelsey Clark, and Donald P. Green, "Prejudice Reduction: Progress and Challenges," *Annual Review of Psychology* 72 (2021): 533–560, https://doi.org/10.1146/annurev-psych-071620-030619; Toni Schmader, Tara C. Dennehy, and

Andrew S. Baron, "Why Antibias Interventions (Need Not) Fail," *Perspectives on Psychological Science* 17 (5) (2022): 1381–1403, https://doi.org/10.1177/17456916211057565; and Anthony G. Greenwald, Nilanjana Dasgupta, John F. Dovidio, et al., "Implicit-Bias Remedies: Treating Discriminatory Bias as a Public-Health Problem," *Psychological Science in the Public Interest* 23 (1) (2022): 7–40, https://doi.org/10.1177/15291006211070781.

[41] Miao K. Qian, Paul C. Quinn, Gail D. Heyman, et al., "Perceptual Individuation Training (But Not Mere Exposure) Reduces Implicit Racial Bias in Preschool Children," *Developmental Psychology* 53 (5) (2017): 845–859, https://doi.org/10.1037/dev0000290; and Antonya M. Gonzalez, Jennifer R. Steele, Evelyn F. Chan, et al., "Developmental Differences in the Malleability of Implicit Racial Bias Following Exposure to Counterstereotypical Exemplars," *Developmental Psychology* 57 (1) (2021): 102–113, https://doi.org/10.1037/dev0001128.

[42] Geraldine Dawson, Sally Rogers, Jeffrey Munson, et al., "Randomized, Controlled Trial of an Intervention for Toddlers with Autism: The Early Start Denver Model," *Pediatrics* 125 (1) (2010): e17–e23, https://doi.org/10.1542/peds.2009-0958; and Naja Ferjan Ramírez, Kaveri K. Sheth, and Patricia K. Kuhl, "The Effects of Age, Dosage, and Poverty on Second Language Learning through SparkLing™ in Infant Education Centers in Madrid, Spain," *International Journal of Environmental Research and Public Health* 18 (23) (2021): 12758, https://doi.org/10.3390/ijerph182312758.

[43] John F. Dovidio, Kerry Kawakami, and Samuel L. Gaertner, "Implicit and Explicit Prejudice and Interracial Interaction," *Journal of Personality and Social Psychology* 82 (1) (2002): 62–68, https://doi.org/10.1037//0022-3514.82.1.62; Jennifer A. Richeson and J. Nicole Shelton, "Thin Slices of Racial Bias," *Journal of Nonverbal Behavior* 29 (1) (2005): 75–86, https://doi.org/10.1007/s10919-004-0890-2; and John F. Dovidio, "Racial Bias, Unspoken But Heard," *Science* 326 (5960) (2009): 1641–1642, https://doi.org/10.1126/science.1184231.

[44] J. Nicky Sullivan, Jennifer L. Eberhardt, and Steven O. Roberts, "Conversations About Race in Black and White U.S. Families: Before and After George Floyd's Death," *Proceedings of the National Academy of Sciences* 118 (38) (2021): e2106366118, https://doi.org/10.1073/pnas.2106366118; Jamie L. Abaied and Silvia P. Perry, "Socialization of Racial Ideology by White Parents," *Cultural Diversity and Ethnic Minority Psychology* 27 (3) (2021): 431–440, https://doi.org/10.1037/cdp0000454; and Leoandra Onnie Rogers, Katharine E. Scott, Finn Wintz, et al., "Exploring Whether and How Black and White Parents Talk with Their Children about Race: M(ai)cro Conversations about Black Lives Matter," *Developmental Psychology* (advance online publication, 2024), https://doi.org/10.1037/dev0001693.

[45] Ibid.

[46] For research on parental beliefs and values about racial socialization, see Amber D. Williams and Meeta Banerjee, "Ethnic-Racial Socialization among Black, Latinx, and White Parents of Elementary School-Age Children," *Journal of Social Issues* 77 (4) (2021): 1037–1062, https://doi.org/10.1111/josi.12493. For research on children's own beliefs about race, self, and society, see Leoandra Onnie Rogers and Andrew N. Meltzoff, "Is Gender More Important and Meaningful Than Race? An Analysis of Racial and Gender Identity among Black, White, and Mixed-Race Children," *Cultural Diversity and Ethnic Minority Psychology* 23 (3) (2017): 323–334, http://doi.org/10.1037/cdp0000125; and Leoandra Onnie Rogers, Ursula Moffitt, and Christina Foo, "'Martin Luther King Fixed It': Children Making Sense of Racial Identity in a Colorblind Society," *Child Development*

92 (5) (2021): 1817–1835, https://doi.org/10.1111/cdev.13628. For research on match/mismatch between parents versus children's beliefs about racial biases, see Katharine E. Scott, Tory L. Ash, Bailey Immel, et al., "Engaging White Parents to Address Their White Children's Racial Biases in the Black-White Context," *Child Development* 94 (1) (2023): 74–92, https://doi.org/10.1111/cdev.13840.

47 Skinner and Meltzoff, "Childhood Experiences and Intergroup Biases Among Children."

48 U.S. Government Accountability Office, *K–12 Education: Student Population Has Significantly Diversified, but Many Schools Remain Divided along Racial, Ethnic, and Economic Lines* (Washington, D.C.: U.S. Government Accountability Office, 2022), https://www.gao.gov/assets/gao-22-104737.pdf; and Erica Greenberg and Tomas Monarrez, *Segregated from the Start: Comparing Segregation in Early Childhood and K–12 Education* (Washington, D.C.: Urban Institute, 2019), https://www.urban.org/features/segregated-start.

49 Urban Institute, *Segregated Neighborhoods, Segregated Schools? More than 60 Years after Brown v. Board of Education, Why Does School Segregation Persist?* (Washington, D.C.: Urban Institute, 2018), https://www.urban.org/data-tools/segregated-neighborhoods-segregated-schools.

50 Jordan G. Starck, Travis Riddle, Stacey Sinclair, and Natasha Warikoo, "Teachers Are People Too: Examining the Racial Bias of Teachers Compared to Other American Adults," *Educational Researcher* 49 (4) (2020): 273–284, https://doi.org/10.3102/0013189X20912758; and Kate M. Turetsky, Stacey Sinclair, Jordan G. Starck, and J. Nicole Shelton, "Beyond Students: How Teacher Psychology Shapes Educational Inequality," *Trends in Cognitive Sciences* 25 (8) (2021): 697–709, https://doi.org/10.1016/j.tics.2021.04.006.

51 Carol Lee, Gregory White, and Dian Dong, eds., *Educating for Civic Reasoning and Discourse* (Washington, D.C.: National Academy of Education, 2021), https://naeducation.org/educating-for-civic-reasoning-and-discourse-report.

Uncovering Implicit Racial Bias in the Brain: The Past, Present & Future

Jennifer T. Kubota

Neuroscience is a fantastic tool for peeking inside our minds and unpacking the component processes that drive social group biases. Brain research is vital for studying racial bias because neuroscientists can investigate these questions without asking people how they think and feel, as some individuals may be unaware or reluctant to report it. For the past twenty-five years, neuroscientists have diligently mapped implicit racial bias's neural foundations. As with any new approach, the emergence of neuroscience in studying implicit racial bias has elicited excitement and skepticism: excitement about connecting social biases to biological machinery, and skepticism that neuroscience may provide little to our understanding of social injustice. In this essay, I dive into what we have learned about implicit racial bias from the brain and the limitations of our current approach. I conclude by discussing what is on the horizon for neuroscience research on racial bias and social injustice.

Racism is embedded in U.S. culture and systems. A foundation built not by accident, but with deliberate determinism, by and upon the enslaved and oppressed to uphold hegemony and hierarchy. Racism was enshrined in the Constitution through a provision limiting African Americans to three-fifths personhood. Years of slavery, lynching, and brutalism were supported by racist legislation, leading to a segregationist and discriminatory society. Despite this scorched foundation, after the U.S. civil rights movement, there was optimism for some that the country was forging a new path with the introduction of normative and legal changes. This optimism was ostensibly supported by national surveys revealing emerging positive sentiment toward Black people.[1] Contrasting the racialized beliefs before the 1960s with the changing culture offered signs of hope, as the nation appeared to support the principles of racial integration and equal treatment openly and enthusiastically. Enter the myth of racial progress, whereby White Americans began to falsely believe that the United States had achieved considerable racial equality, when in fact racial disparities were (and are) deeply ingrained in American society.[2] This myth was coupled with the growing societal perspective that bias, discrimination, and racism were wrong, and expressing such bias was, in many spheres of society, frowned upon.[3]

https://doi.org/10.1162/daed_a_02050

The Past: The Origins of Implicit Bias

Were racial biases actually decreasing? And could scientists find a way to assess the tension between a cultural shift toward favoring equality and the reality of racial bias embedded in systems and apparent in daily life? Intergroup scholars at the time thought that public opinion surveys may not accurately capture people's true beliefs. Furthermore, although there was progress in legal changes and norms, the racist structures and systems, prevalent stereotypes and prejudices, and human motivation to favor one's group cast looming shadows on equality. For example, although public opinion polls in the 1970s and 1980s showed the country moving away from explicit racial biases, discrimination could be identified in laboratory experiments, particularly when the participants were unaware that racial bias was the experiment's focus. Social psychologists Faye Crosby, Stephanie Bromley, and Leonard Saxe summarized these studies and called into question the assumption that verbal reports accurately reflect individuals' sentiments.[4] They concluded that White Americans were more prejudiced than they were willing to admit, theorizing that individuals might not disclose their genuine opinions on surveys for fear of judgment, but would reveal them when they felt safe or were unaware that researchers were investigating racial bias. Researchers at the time believed that even individuals who valued equality would sometimes exhibit discriminatory behavior.[5] Consequently, many considered self-reports to be unreliable. This perspective was consistent with a broader trend in social psychology that approached self-reports with skepticism and favored cognitive tasks as a more reliable measure of attitudes.[6] At this time, cognitive psychology was exploring how priming a concept for a subject (such as by showing someone a word or picture) before they performed a given task could shape their responses. This general approach, that one can prime a concept that activates related concepts or prepares folks to view others in a related way without needing self-reflection, would significantly shape the development of implicit racial bias measures.

Researchers viewed behavioral implicit measures as a way to understand why individuals who consciously reject prejudice, such as egalitarians, still exhibit biased behavior. Enter Patricia Devine. In 1989, Devine, a social psychologist specializing in prejudice and stereotypes, suggested that discriminatory behavior and self-reports represented authentic psychological processes in conflict.[7] One process was automatic antipathy, resulting from repeated exposure to negative cultural information about social groups. The other was a more deliberate reflection of genuine beliefs or values (for example, I want to be or should be egalitarian). This idea fostered the modern perspective that stereotypes and prejudices are learned associations influenced partly from culture.[8] Devine's perspective was popular among researchers: it offered optimism (people might be able to control their bias), intervention possibilities (perhaps we can foster self-control of bias),

and historical resonance (this is why self-reports are deviating from widespread systemic biases in wealth, health, education, policing, and employment). However, Devine did not provide a direct measure of spontaneous group associations (that is, implicit bias) but instead attempted to demonstrate them using "unobtrusive" methods: namely implicit behavioral measures.[9]

Social psychologist Russell H. Fazio and colleagues, and later Anthony G. Greenwald and colleagues, introduced indirect behavioral measures of spontaneous group associations, known as implicit bias, and introduced the term "implicit social cognition" to describe cognitive processes related to social psychological constructs that occur outside of awareness or control.[10] The general premise is that people lack self-reflective access to the cause of their behavior and are terrible at introspection, and that these new measures of implicit bias avoided the need for accurate self-reflection.[11] Researchers could immediately see the appeal of tapping into biases without needing self-report. Over the next twenty-five years, there was a proliferation of implicit behavioral measures and the application of these measures to real-world domains, such as mental health, consumer decision-making, policing, legal decisions, education, health care, and political behavior.[12] The popularity of these measures only gained as time went on. Implicit bias, measured behaviorally, quickly entered the public lexicon and was even mentioned in the 2020 presidential debate between Hillary Clinton and Donald Trump.[13]

But what is implicit bias, and how is it measured? Intergroup implicit (and explicit) associations are evaluations or beliefs about social groups. One difference between these associations is that individuals report explicit evaluations, whereas implicit associations are measured indirectly.[14] Therefore, like other memory/evaluative associations, implicit race-based evaluations are partly acquired through repeated paired associations with a group (such as culturally or environmentally learned association) and can be applied without deliberation. Explicit attitudes are also partly acquired through environmental/cultural learning. Therefore, discriminatory responses can occur without intention, even when counter to deliberative unbiased beliefs. Individuals may thus feel genuine positivity about an out-group (that is, a member from a different racial group than the perceiver) and support equality, but still exhibit implicit bias.[15]

Over the last twenty-four years, researchers have consistently found that the majority of people in the United States show some degree of negative implicit associations about Black people and positive associations about White people. Our research has even observed these associations with self-identified Black Americans when they interact more with White people.[16] Furthermore, researchers have observed greater implicit bias in more segregated counties in the United States, in places with a history of chattel slavery, or among individuals whose parents have greater implicit racial bias.[17] Therefore, substantial evidence shows that the systems, culture, and whom we interact with shape implicit racial biases. People are

absorbing these associations about groups from their environments whether they want to or not – even negative associations about their own group.

As the implicit bias revolution gained steam in social psychology, researchers wondered whether there were ways to assess the evaluative and cognitive processes underlying implicit bias without a response requirement. At this point, most of the research was conducted by social psychologists, and they were rightfully concerned that individuals would attempt to control their behaviors to avoid appearing biased when forced to respond or were aware that the measure might assess racial bias. Moreover, researchers were concerned that some implicit bias measures were potentially contaminated by task demands (such as forcing people to compare groups or to make a response).[18] These two factors partly motivated a new era in implicit bias research in which scholars sought means to assess racial bias uncontained by these factors. Starting in the 1990s and increasing in the early 2000s, a new field took form, social neuroscience, that allowed researchers to investigate implicit racial bias via neural measures without asking people what they think, while also allowing scholars to outline the underlying levers and gears that produce these biases.

The Present: Social Neuroscience of Implicit Bias

Neuroscience methods allow researchers to assess implicit processes impacting how we think, feel, and behave toward marginalized/minoritized individuals in real time without needing self-report or behavioral responses. In these ways, neuroscience is a fantastic tool for peeking inside our minds and unpacking component processes contributing to behavior, allowing scholars to understand how the brain works at the cellular and molecular levels, how different brain regions are connected and interact, and how information is processed and integrated. Fundamentally neuroscience allows us to measure mechanisms (think of mechanism as what is under the hood making the car move; the *how* of implicit bias). Knowledge about mechanisms can shed light on underlying cognitive, social, emotional, and behavioral processes. Because of this, neuroscience provides valuable insights into the cognitive and affective processes that drive racial bias, and minimizes many of the criticisms of behavioral measures of implicit bias.

The use of neuroscience in social psychology is a relatively recent development that has gained momentum.[19] More concretely, neuroimaging has provided several advantages for studying racial bias, including assessing ongoing psychological processes without the intrusive questions and socially desirable responses that can occur with self-report.[20] Moreover, neuroimaging offers sensitivity to the engagement of distinct psychological processes that underlie otherwise similar behavior, allowing scholars to determine, for example, whether lapses in cognitive control rather than negative evaluations are more predictive of implicit bias.[21] As-

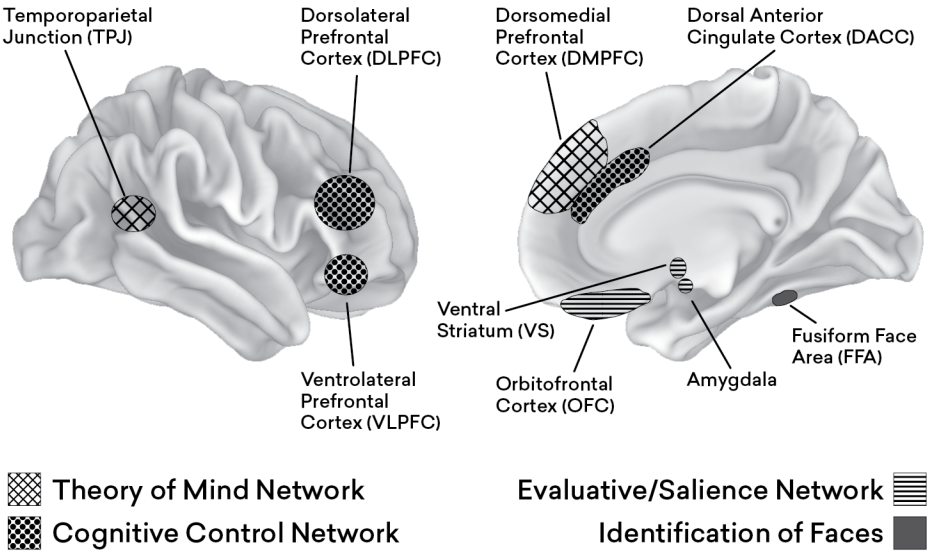
sessing simultaneously multiple and rapid unfolding processing is extremely difficult, if not impossible, with most implicit behavioral measures.

In 1992, social psychologists John Cacioppo and Gary Berntson introduced the term social neuroscience to describe an interdisciplinary approach to mapping social behavior and cognition by integrating our understanding of psychology with neuroscience (that is, the mind and body). From there, the field rapidly expanded due to the increased availability of noninvasive central nervous system measures. Among the first technique was event-related brain potentials (ERPs), which are derived from electroencephalograms (EEGs) and measure electrical activity of the brain at the scalp.[22] ERPs allow scholars to assess electrical activity in real-time as people view others or respond to prompts. ERPs are particularly useful for studying racial bias because they allow researchers to understand when a process is happening in time. Researchers often strip away the context in the lab and simply show people faces varying in social group membership. When they do so, they find that individuals process information about perceived race, gender, age, status, and emotion within two hundred milliseconds of encountering someone.[23] That is incredibly fast! This tells us that information about these social categories gets into our minds early and can guide impressions. Most important, it is spontaneous. Even when we ask people to stop, the brain still processes social category information rapidly.[24]

We can not only understand when things are happening in time with EEG but also view which areas of the brain are processing social group information using functional magnetic resonance imaging (fMRI), which measures changes in blood flow in brain regions while participants perform tasks or view images, helping us map mechanisms, and is another essential tool for neuroscientists.[25] Although it does not provide precise timing information like ERPs, it offers excellent spatial resolution by providing information about the specific brain areas associated with mental operations. Over the years, researchers have found a host of regions involved in social group processing, including a few usual neural suspects.[26] Specifically, these regions support the identification of faces (fusiform face area [FFA]), the evaluation of others based on their perceived race (orbitofrontal cortex [OFC], amygdala, and ventral striatum [VS]), how we represent the minds of others based on their perceived race or perform theory of mind (dorsomedial prefrontal cortex [DMPFC] and temporoparietal junction [TPJ]), and the regulation of bias (dorsolateral prefrontal cortex [DLPFC], anterior cingulate cortex [ACC], and ventrolateral prefrontal cortex [VLPFC]). Importantly, these areas are very similar to the areas that are involved in the processing and regulation of other emotional and social stimuli more generally, as seen in Figure 1.

Therefore, researchers have found that individuals process perceived race both *extremely* fast and in the same way they process other emotion-laden stimuli. Most important, it is unintentional. Even when we ask people to stop, the brain

Figure 1
Brain Regions Supporting Processing of Social Group Membership



The brain regions supporting processing of social group membership include the Identification of Faces (FFA), Evaluative/Salience Network (amygdala, VS, and OFC), Theory of Mind Network (TPJ and DMPFC), and Cognitive Control Network (DLPFC, VLPFC, and DACC). Source: Illustration by the author.

still processes this information rapidly. During this time, neuroscientists began to disentangle the processing of another's perceived race from the production of implicit bias.[27] Although perceiving race, whether accurate or not, is necessary to produce implicit associations – that is, one must categorize someone as belonging to a group to bring to mind (or activate) stereotypes and prejudices about the group – it is not sufficient. Just because folks in the United States process race does not mean they will have implicit or explicit biases, or that perceived race is an innate category. Race is a culturally and socially constructed category imbued with evaluative and semantic meaning. Consequently, the brain processes this culturally and socially constructed category similarly to other emotionally charged or salient information in the environment that culture or one's social network has deemed positive, negative, or important.

For the past twenty-five years, neuroscientists have diligently mapped implicit racial bias's neural foundations. One key finding is that implicit racial bias ap-

pears to be rooted partly in the brain's evaluative system, which can operate spontaneously. This is not surprising given that individuals learn evaluative associations about groups from the culture and their environment, and this learning is then reflected in patterns of brain activity. One region that is part of the evaluative brain network, the amygdala, has been a common focus of fMRI studies examining how people process perceived race.[28] The amygdala is vital for retaining, forming, and expressing negative evaluations, including fear.[29] Additionally, the amygdala has a more extensive function in quickly identifying biologically significant stimuli, facilitating rapid attention and memory.[30] Specifically, the amygdala responds with greater activation to faces of people from racial groups that are less familiar or positively viewed.[31] Differences in amygdala activity to perceived race have sometimes, though certainly not always, correlated with implicit racial bias (and typically not explicit bias).[32] Despite some disagreement on its interpretation, the general consensus is that the discovery of amygdala responses to perceived race in the U.S. context suggests that White individuals perceive Black people as highly noticeable (salient) and potentially threatening.

Recent research has uncovered significant variation in how the amygdala responds to perceived race and the degree to which the amygdala is solely or partly producing implicit bias (if even producing it at all).[33] For example, when additional information about group membership or traits is available, preferential amygdala activity is frequently absent based on perceived race.[34] Therefore, our comprehension of the amygdala's role in perceived race-based assessment is more intricate and adaptable than previously believed. The current agreement suggests that this region is not the primary source of implicit racial bias. Instead, amygdala sensitivity to perceived race might result from several factors, ranging from culturally learned stereotypes to the social threat of being seen as prejudiced.[35] These cultural associations can differ from one person to another based on their formative experiences. In line with this perspective, more interracial interactions during childhood correlates with decreased amygdala responses to familiar (compared to unfamiliar) perceived Black individuals.[36]

Another key finding is that implicit bias depends on self-regulation. The ability to adjust and control our behavior is a valuable human skill that enables us to act flexibly to achieve goals. Researchers, such as Devine, have suggested that individuals in the United States have conflicts between culturally or environmentally acquired stereotypes and prejudices and personal or societal norms to appear antiracist.[37] In other words, inconsistency arises from the desire to respond without racial prejudice and the activation of stereotypical or prejudicial associations. This has led researchers to suggest that implicit bias is partly a self-control failure or an inability to regulate that conflict. Existing research has shed light on the neural mechanisms underlying this self-regulation process, revealing a network of brain regions that are believed to identify the need for control, maintain regulatory

goals, and facilitate the selection of actions that align with the desired goals while inhibiting actions that do not (that is, goal-congruent versus goal-incongruent responses).

The anterior cingulate cortex was initially linked to detecting conflicts between prepotent and intentional response tendencies.[38] However, in recent years, new models of ACC function suggest that this region is involved in computing the value of engaging in cognitive control based on various factors, including task difficulty, feedback, uncertainty, and reward.[39] A U.S. study found that implicit racial bias increases ACC activity when viewing perceived Black individuals (as opposed to White individuals) when the faces are less prototypical (that is, inconsistent with racial stereotypes).[40] These studies assume that cognitive conflict arises due to differences in the participants' implicit biases and their motivations to be and/or appear egalitarian. Sensitivity to race in the ACC and other control-related brain areas is consistently most evident when folks know the study is about race and when participants believe that task responses indicate racial bias.[41] Research also finds that a greater internal drive to respond without prejudice may amplify cognitive conflict, even without explicit instructions to control racial bias.[42] Therefore, when people are cognizant that racial bias might be assessed, they may engage in more self-control as a strategy to avoid bias.

Another important region involved in self-regulation of racial bias is the dorsolateral prefrontal cortex (DLPFC).[43] The DLPFC is responsible for the executive control of sensory and motor operations that align with operational goals.[44] Recent research indicates that younger adults exhibit greater DLPFC activity when viewing perceived Black faces compared to perceived White faces than do older adults, who, in this study, had less self-regulation abilities.[45] The DLPFC and ACC may work together to regulate implicit racial bias.[46] The ACC may detect conflicts between explicit intentions and implicit associations, while the DLPFC may help to regulate the expression of implicit bias.[47] However, like the amygdala, there is less evidence that self-regulation, as reflected by DLPFC or ACC activity, is the driver of implicit bias alone.

The work seems to suggest that the evaluative brain network and the cognitive-control regulatory brain network both seem to partly contribute to implicit racial bias. Therefore, current research suggests that implicit racial bias is a complex phenomenon involving multiple neural pathways and mechanisms that rely on evaluative and cognitive control systems. While neuroscience can help us understand the underlying processes, it is essential to again underscore that implicit bias is not just a matter of individual brain activity but also a product of cultural and social factors that shape our biases. Research suggests that vital brain systems have been co-opted, in a sense, to process this socially constructed category – race – and help to produce implicit bias because the culture has imbued racial groups with meaning, particularly negative biases toward marginalized or minoritized groups. Just

because researchers can identify how the brain processes others based on race does not mean racial bias is innate.

Because culture and the environment have amplified biases toward marginalized or minoritized groups, intervening at the systemic level would likely have the most significant impact. However, neuroscience can inform how changes in our environment or new pieces of information shape implicit bias, providing valuable insights about the flexibility of these processes. One of the most promising avenues for reducing racial bias (both implicit and explicit) that has behavioral and neuroscience support is via interracial contact. Psychologist Jasmin Cloutier and colleagues were some of the first neuroscientists to investigate how contact influences neural mechanisms and reduction in racial bias.[48] In their 2011 study, people were first familiarized with perceived Black faces and White faces, then went into the scanner and viewed faces, some of which were new, and some were the same faces they had already seen. What they found for the novel faces (faces they had never seen) was similar to what neuroscientist Elizabeth Phelps and colleagues detected in 2000.[49] For self-identified White Americans, the amygdala responses were greater to perceived Black individuals than White individuals. However, the amygdala difference disappeared when respondents were more familiar with the perceived individuals. This was even more pronounced for folks with more childhood interracial contact. In fact, Jasmin Cloutier and colleagues found in their 2014 study that greater interracial childhood contact reduced amygdala responses in adulthood eighteen-plus years later.[50] Around the same time, neuroscientist Eva Telzer and colleagues found that increased early deprivation, characterized by a delayed age of adoption, correlated with heightened amygdala differences toward race. These findings highlight the influence of early social intergroup interactions on the functioning of the amygdala in later stages of life.[51]

Interracial contact also shapes how individuals mentalize about out-group members. The ability to mentalize, also known as "theory of mind," enables humans to make inferences about the emotions, intentions, goals, and motivations of others, thereby aiding in navigating complex social interactions. One great thing about neuroscience tools is that they allow scientists to measure mentalizing in real time. Research indicates that the dorsomedial prefrontal cortex (DMPFC) and temporoparietal junction (TPJ), among other brain regions, are consistently activated when individuals infer the mental states of others, particularly for in-group members relative to out-group members.[52] However, recent research suggests that folks with more interracial contact (for example, quality contact with Black individuals for White participants) engage in similar mentalizing when viewing perceived Black faces and White faces.[53] Moreover, mentalizing processes may help perceivers determine whether they observe social injustice during violent interracial interactions. For example, our recent research with self-identified White Americans finds that greater interracial contact increases men-

talizing when watching videos showing violent arrests of perceived Black civilians by White officers.[54] Together this work points to the importance of mentalizing processes in diminishing racial bias and facilitating the identification of racial injustice. It appears that mentalizing may act on explicit rather than implicit bias, but more research must be done to investigate this possibility.

Early fMRI work focused on specific brain regions, but contemporary neuroscience considers how entire brain networks coordinate when encountering or interacting with others. What is fascinating is that interracial contact not only determines how one region of the brain responds – for example, the amygdala – but our recent research demonstrates that contact shapes how entire brain networks respond to others, particularly those involved in social evaluation and mentalizing.[55] Therefore, contact has a powerful impact on how our brain works in concert when encountering others. This research, combined with excellent behavioral work in social psychology, suggests that intergroup contact may work as an intervention in some situations, but it is only sometimes feasible. It can put marginalized and minoritized folks in spaces they might not want to be in, and creating meaningful contact where strangers build relationships is a challenge. So, it is not a perfect solution.

Overall, neuroscience can provide valuable insights into the evaluative and cognitive mechanisms underlying implicit bias and the effects of different interventions and social contexts on these biases. Additionally, the new network-neuroscience approach may be more suited for mapping not only the constellation of factors that give rise to implicit bias but also how they function in concert and how changes in the coordination of these networks may reduce implicit bias. By incorporating insights from neuroscience into implicit bias research, we may better understand how implicit biases operate and identify effective strategies for reducing their impact.

The Future: Skepticism and What Is on the Horizon for Neuroscience Research of Implicit Bias

Although neuroscience and social psychology have provided essential insights into implicit bias's origins, production, and consequences, the field of implicit bias has faced criticism. For one, researchers need to clarify how crucial implicit bias is in producing everyday discrimination.[56] Moreover, implicit bias training can enhance knowledge on the topic but does not consistently reduce implicit bias or impact behavior.[57] For example, while many individual studies have shown significant relationships between implicit measures and discriminatory behaviors, the overall impact tends to be small.[58] This has led some critics to consider the construct insignificant to our understanding of discrimination or racism.[59] Although

this possibility is important, it is premature to write off implicit bias entirely. For one, it is still vital to understand every contributing factor to racial bias and racism. Moreover, different implicit measures show different predictive validity, so throwing them all out rather than understanding their strengths and weaknesses could impair our understanding.[60] Finally, it appears that implicit bias at the population level is a relatively good predictor of some aspects of systemic biases and racism, and some neuroscientists have started to map how implicit bias at the population level shapes neural responses.[61] After all, individuals make up systems and institutions. Individual biases and racialized interactions are ingrained into institutional policies and societal systems, propagating the development and perpetuation of systemic racism.

However, during one-on-one interactions, it appears that having implicit racial bias does not necessarily indicate the presence of a single person's racial prejudice or the likelihood that someone will discriminate, as going from associations to actions is complex and multifaceted. Instead, the person, situation, and culture influence discriminatory actions. It remains unclear how critical implicit bias is to structural racism over and above needs for power or status, in-group, the group one identifies with or belongs to, favoritism, or explicit bias. Therefore, it may be inappropriate to generalize from a single implicit bias behavioral or neural measure, even if it pertains to a significant conceptual grouping, as it may not reflect a fundamental or widespread change in the level of prejudice in the population or decrease racism. These interpretations must be cautiously approached since social phenomena may continue to be influenced strongly by racism even as implicit bias decreases in the population.[62]

Neuroscience alone cannot fully explain social group biases. Racial bias is shaped by a complex interplay of cognitive, affective, social, cultural, and environmental factors. Neuroscientists can only partially understand this phenomenon as the current methods often focus on one person's mental operations. Although it can provide us with rich information about the mechanisms that occur when we process others from different racial groups, produce implicit bias, or take discriminatory actions, the field is relatively new. There is still much to discover! As we delve deeper into social neuroscience, we must be cautious and mindful of the potential pitfalls that can affect the rigor and inference of neuroscience research, especially when dealing with complex social interactions.[63] For example, most neuroscience research examining how people perceive race and respond to racial out-group members typically shows pictures of faces that are disembodied and out of context. Although this allows researchers to isolate different aspects of the process, it does not represent the multitude of information and contexts available in real-life encounters. Unfortunately, these factors may be critical drivers or mitigators of bias, but without investigating them, we may have a blind spot. Moreover, most current research fails to examine whether neural processes

predict discriminatory behavior. In other words, just because we see an area of the brain involved in processing individuals of different perceived racial groups, it does not mean that part of the brain is necessary for discrimination. Therefore, we do not have a sense of the predictive power of neuroscience for understanding real-world discrimination.[64]

Using brain imaging techniques to study implicit racial bias has been criticized for potentially reinforcing the idea of inherent or innate racial bias rather than focusing on the social and cultural factors contributing to biased attitudes and behaviors. The public and even other scholars will misconstrue a response in the brain as evidence of the innateness of bias. Social neuroscientists have firmly pushed back against this interpretation, suggesting that culture largely drives these biases, but this misinterpretation still plagues the science.[65] Moreover, once neuro measures are involved, the tendency to view the process as innate almost medicalizes the solutions. For example, one 2012 study demonstrated that a drug, a common beta blocker propranolol, reduced implicit bias.[66] One can imagine the headlines: Pills to Cure Racism! While potentially providing insights into some biological processes, this study raised troubling public discussions about developing a drug to treat racism and, in effect, biologizing racism.[67] Others suggested that focusing on a particular brain region and levying a neurological intervention would cure this social ill. The truth is that brain regions, like our neurophysiology and endocrinology, are intertwined, and each typically has multiple functions. In this way, these statements are wildly inappropriate and highly inaccurate, representing extreme forms of how neuroscience research can be misinterpreted. There is no magic pill. There is no neurological or biochemical solution, and making these claims distracts from the historical and social factors that shape and reinforce racism. Racism is rooted in our structures and systems. How we process information is a byproduct of those systems. Neuroscience measures allow us to assess that byproduct with more nuance than behavioral measures alone. They can guide our understanding of racial bias, but we cannot and should not turn to a biological solution for racism.

In addition, despite the inherent dynamism of social interactions and processes, there is a lack of neuroscience work examining dynamic intergroup interactions. New techniques are now changing this. To increase the generalizability of brain research, scholars have adopted approaches such as hyperscanning, mobile EEG, fNIRs (functional near-infrared spectroscopy), and portable physiological tools, which enable us to extend our inquiry to real dynamic interactions and reach communities that were previously difficult to include due to financial or geographical constraints. These portable methodologies also remove cost barriers associated with fMRI and expand the sample and researcher demographics who can participate in social neuroscience. Ultimately, this will improve our understanding of neural correlates of racial bias because we can assess these biases

during real interactions and with samples of individuals other than undergraduates at universities rich enough to afford an fMRI scanner.

While neuroscience research on implicit bias has provided essential insights that even behavioral research alone could not provide (for example, the role of mentalizing in intergroup bias), it is just a starting place for much-needed research. Current racial bias research may not generalize across samples, stimuli, cultures, or historical points. This is vital because race is a cultural construct, with the meaning changing across history and cultures. Most of the current research focuses on White folks in the United States viewing perceived Black faces and White faces. Additionally, the people included in the studies (the sample) and whom they view (the stimuli) are typically young and self-identify as cisgender men or women. Moreover, researchers often do not even ask about political ideology or sexual orientation. These oversights impair our understanding as certain groups are more or less likely to attend to and discriminate against others based on perceived race. Therefore, we know little about how intersectional identities shape how people process race, representing a more naturalistic understanding of intergroup dynamics.

Social psychologists and social neuroscientists have primarily examined bias with people who espouse equality. However, plenty of folks explicitly hate others based on their social group of belonging. This is a critical missing piece in our understanding of racial bias as these individuals express hate and an intention to act upon it. They might like intergroup discrimination and violence and perceive it as just. Understanding the drivers of explicit bias with neuroscience and behavioral research methods (not simply implicit bias) could allow researchers to characterize who is vulnerable to espousing hate or joining hate groups, what processes underlie explicit bias, and how we may intervene when individuals are entrenched in hate.

By examining the human brain, both neuroscience and the study of implicit bias can provide insight into why we treat others with cruelty or kindness and exhibit empathy or apathy. While social neuroscience has yet to contribute significantly to our understanding of overall social injustice, the discipline is poised to push this frontier further. However, achieving social justice requires understanding the complex issues, including historical and structural factors, that affect equity and inclusion and mitigate racial bias, and this understanding must be integrated into our scholarship. Although neuroscience can uncover our biases and prevent us from denying the inclinations of our minds, it does not justify maintaining or acting on those biases. By mapping how our brains function, we can acknowledge and start to understand racial biases. This awareness may assist us in defeating these biases in everyday interactions and collaborating toward a more fair and equitable society. To do so, we must consider structures, individuals, and groups in our research and be inclusive in our scientific endeavors. Final-

ly, addressing implicit bias alone is insufficient to create a genuinely united society in the twenty-first century. The most effective means of changing bias is likely through altering the overall social structures and conditions that underpin and reinforce racism. A united national leadership and culture must speak out against racial bias, discrimination, poverty, failing health care and schools, and other insidious factors contributing to injustice. The neuroscience of implicit bias must be understood as a situated approach, whereby we recognize the significance of environmental and cultural factors in shaping the cognitive and evaluative mechanisms that give rise to racial bias.

ABOUT THE AUTHOR

Jennifer T. Kubota is a Senior Ford Fellow and an Associate Professor in the Departments of Psychological and Brain Sciences and Political Science and International Relations at the University of Delaware. She codirects the Impression Formation Social Neuroscience Lab and is the Director for Equity and Inclusion in the Department of Psychological and Brain Sciences. She has published over forty articles in journals such as *Nature Neuroscience*, *Nature Human Behaviour*, *Psychological Science*, *Perspectives in Psychological Science*, *Journal of Experimental Social Psychology*, and *Biological Psychology*. She was awarded a Senior Ford Fellowship in 2022 and is a Fellow of the Society for Experimental Social Psychology. She is on the governing boards of the Social and Affective Neuroscience Society and the SPARK Society. She is also the Chair of the Equity, Diversity, Inclusion, and Justice Committee for the Social and Affective Neuroscience Society, an Associate Editor at the *Journal of Personality and Social Psychology*, and an Editorial Board member at *Scientific Reports*.

ENDNOTES

[1] Lawrence Bobo, "Racial Attitudes and Relations at the Close of the Twentieth Century," in *America Becoming: Racial Trends and Their Consequences*, ed. Neil J. Smelser, William Julius Wilson, and Faith Mitchell (Washington, D.C.: National Academies Press, 2001), 264–301.

[2] Jennifer A. Richeson, "Americans Are Determined to Believe in Black Progress," *The Atlantic*, September 2020, https://www.theatlantic.com/magazine/archive/2020/09/the-mythology-of-racial-progress/614173; and Ivuoma N. Onyeador, Natalie M. Daumeyer, Julian M. Rucker, et al., "Disrupting Beliefs in Racial Progress: Reminders of Persistent Racism Alter Perceptions of Past, but Not Current, Racial Economic Equality," *Personality and Social Psychology Bulletin* 47 (5) (2021): 753–765.

[3] John F. Dovidio and Samuel L. Gaertner, "Aversive Racism," *Advances in Experimental Social Psychology* 36 (2004): 1–52.

[4] Faye Crosby, Stephanie Bromley, and Leonard Saxe, "Recent Unobtrusive Studies of Black and White Discrimination and Prejudice : A Literature Review," *Psychological Bulletin* 87 (3) (1980) : 546–563.

[5] Ibid.

[6] Robert S. Wyer and Donal E. Carlston, *Social Cognition, Inference, and Attribution* (London : Psychology Press, 1979).

[7] Crosby, Bromley, and Saxe, "Recent Unobtrusive Studies of Black and White Discrimination and Prejudice."

[8] Benedek Kurdi, Alison E. Seitchik, Jordan R. Axt, et al., "Relationship between the Implicit Association Test and Intergroup Behavior : A Meta-Analysis," *American Psychologist* 74 (5) (2019) : 569 ; and Tessa E. S. Charlesworth, Aylin Caliskan, and Mahzarin R. Banaji, "Historical Representations of Social Groups across 200 Years of Word Embeddings from Google Books," *Proceedings of the National Academy of Sciences* 119 (28) (2022) : e2121798119.

[9] Thomas K. Srull and Robert S. Wyer Jr., "The Role of Category Accessibility in the Interpretation of Information about Persons : Some Determinants and Implications," *Journal of Personality and Social Psychology* 37 (10) (1979) : 1660–1672.

[10] Russell H. Fazio, Joni R. Jackson, Bridget C. Dunton, and Carol J. Williams, "Variability in Automatic Activation as an Unobtrusive Measure of Racial Attitudes : A Bona Fide Pipeline ?" *Journal of Personality and Social Psychology* 69 (6) (1995) : 1013–1027 ; Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz, "Measuring Individual Differences in Implicit Cognition : The Implicit Association Test," *Journal of Personality and Social Psychology* 74 (6) (1998) : 1464–1480 ; and Anthony G. Greenwald and Mahzarin R. Banaji, "Implicit Social Cognition : Attitudes, Self-Esteem, and Stereotypes," *Psychological Review* 102 (1) (1995) : 4–27.

[11] Richard E. Nisbett and Timothy D. Wilson, "Telling More Than We Can Know : Verbal Reports on Mental Processes," *Psychological Review* 84 (3) (1977) : 231–259.

[12] Matthew K. Nock, Jennifer M. Park, Christine T. Finn, et al., "Measuring the Suicidal Mind : Implicit Cognition Predicts Suicidal Behavior," *Psychological Science* 21 (4) (2010) : 511–517 ; Bethany A. Teachman, Elise M. Clerkin, William A. Cunningham, et al., "Implicit Cognition and Psychopathology : Looking Back and Looking Forward," *Annual Review of Clinical Psychology* 15 (2019) : 123–148 ; Dominika Maison, Anthony G. Greenwald, and Ralph H. Bruin, "Predictive Validity of the Implicit Association Test in Studies of Brands, Consumer Attitudes, and Behavior," *Journal of Consumer Psychology* 14 (4) (2004) : 405–415 ; Andrew Perkins and Mark Forehand, "Implicit Social Cognition and Indirect Measures in Consumer Behavior," in *Handbook of Implicit Social Cognition : Measurement, Theory, and Applications*, ed. Bertram Gawronski and B. Keith Payne (New York : The Guilford Press, 2010), 535–547 ; Calvin K. Lai and Jaclyn A. Lisnek, "The Impact of Implicit-Bias-Oriented Diversity Training on Police Officers' Beliefs, Motivations, and Actions," *Psychological Science* 34 (4) (2023) : 424–434 ; Joshua Correll, Sean M. Hudson, Steffanie Guillermo, and Debbie S. Ma, "The Police Officer's Dilemma : A Decade of Research on Racial Bias in the Decision to Shoot," *Social and Personality Psychology Compass* 8 (5) (2014) : 201–213 ; Kristin A. Lane, Jerry Kang, and Mahzarin R. Banaji, "Implicit Social Cognition and Law," *Annual Review of Law and Social Science* 3 (1) (2007) : 427–451 ; Justin D. Levinson, Robert J. Smith, and Koichi Hioki, "Race and Retribution : An Empirical Study of Implicit Bias and Punishment in America," *UC Davis Law Review* 53

(2) (2019): 839; Jerry Kang and Kristin Lane, "Seeing through Colorblindness: Implicit Bias and the Law," *UCLA Law Review* 58 (2) (2010): 465; Jennifer Neitzel, "Research to Practice: Understanding the Role of Implicit Bias in Early Childhood Disciplinary Practices," *Journal of Early Childhood Teacher Education* 39 (3) (2018): 232–242; Colin A. Zestcott, Irene V. Blair, and Jeff Stone, "Examining the Presence, Consequences, and Reduction of Implicit Bias in Health Care: A Narrative Review," *Group Processes & Intergroup Relations* 19 (4) (2016): 528–542; Luciano Arcuri, Luigi Castelli, Silvia Galdi, et al., "Predicting the Vote: Implicit Attitudes as Predictors of the Future Behavior of Decided and Undecided Voters," *Political Psychology* 29 (3) (2008): 369–387; Anthony G. Greenwald, et al., "Implicit Race Attitudes Predicted Vote in the 2008 U.S. Presidential Election," *Analyses of Social Issues and Public Policy* 9 (1) (2009): 241–253; Bertram Gawronski, Jan De Houwer, and Jeffrey W. Sherman, "Twenty-Five Years of Research Using Implicit Measures," *Social Cognition* 38 (2020): S1–S25; and Bertram Gawronski and Adam Hahn, "Implicit Measures: Procedures, Use, and Interpretation," in *Measurement in Social Psychology*, ed. Hart Blanton, Jessica M. LaCroix, and Gregory D. Webster (London: Routledge, 2018), 29–55.

[13] "Clinton on Implicit Bias in Policing," *The Washington Post*, September 26, 2016, https://www.washingtonpost.com/video/politics/clinton-on-implicit-bias-in-policing/2016/09/26/46e1e88c-8441-11e6-b57d-dd49277af02f_video.html.

[14] Adam Hahn, Charles M. Judd, Holen K. Hirsch, and Irene V. Blair, "Awareness of Implicit Attitudes," *Journal of Experimental Psychology: General* 143 (3) (2014): 1369–1392.

[15] Wilhelm Hofmann, Bertram Gawronski, Tobias Gschwendtner, et al., "A Meta-Analysis on the Correlation between the Implicit Association Test and Explicit Self-Report Measures," *Personality and Social Psychology Bulletin* 31 (10) (2005): 1369–1385.

[16] Brian A. Nosek, Anthony G. Greenwald, and Mahzarin R. Banaji, "The Implicit Association Test at Age 7: A Methodological and Conceptual Review," in *Social Psychology and the Unconscious: The Automaticity of Higher Mental Processes*, ed. John A. Bargh (Abingdon-on-Thames: Routledge/Psychology Press, 2007), 265–292; Anthony G. Greenwald, T. Andrew Poehlman, Eric Luis Uhlmann, and Mahzarin R. Banaji, "Understanding and Using the Implicit Association Test: III. Meta-Analysis of Predictive Validity," *Journal of Personality and Social Psychology* 97 (1) (2009): 17–41; and Jennifer T. Kubota, Jaelyn Peiso, Kori Marcum, and Jasmin Cloutier, "Intergroup Contact throughout the Lifespan Modulates Implicit Racial Biases across Perceivers' Racial Group," *PLOS ONE* 12 (7) (2017): e0180440.

[17] James R. Rae, Anna-Kaisa Newheiser, and Kristina R. Olson, "Exposure to Racial Out-Groups and Implicit Race Bias in the United States," *Social Psychological and Personality Science* 6 (5) (2015): 535–543; B. Keith Payne, Heidi A. Vuletich, and Jazmin L. Brown-Iannuzzi, "Historical Roots of Implicit Bias in Slavery," *Proceedings of the National Academy of Sciences* 116 (24) (2019): 11693–11698; and Sylvia P. Perry, Deborah Wu, Jamie Abaied, et al., "White U.S. Parents' Racial Socialization Messages during a Lab-Based Discussion Task Predict Declines in Their White Children's Pro-White Biases," PsyArXiv, last edited November 27, 2023, https://doi.org/10.31234/osf.io/gf2zx.

[18] Bertram Gawronski, Alison Ledgerwood, and Paul W. Eastwick, "Implicit Bias ≠ Bias on Implicit Measures," *Psychological Inquiry* 33 (3) (2022): 139–155.

[19] Tiffany A. Ito and Jennifer T. Kubota, "Bioelectrical Echoes from a Career at the Cutting Edge: John Cacioppo's Legacy and the Use of ERPs in Social Psychology," *Social Neuroscience* 16 (1) (2021): 83–91; and Tiffany A. Ito and Jennifer T. Kubota, "The Social Neu-

roscience of Social Cognition," in *Handbook of Social Cognition* (Oxford: Oxford University Press, forthcoming).

[20] David M. Amodio, "Can Neuroscience Advance Social Psychological Theory? Social Neuroscience for the Behavioral Social Psychologist," *Social Cognition* 28 (6) (2010): 695–716; Elliot T. Berkman, William A. Cunningham, and Matthew D. Lieberman, "Research Methods in Social and Affective Neuroscience," in *Handbook of Research Methods in Social and Personality Psychology*, ed. Harry T. Reis and Charles M. Judd (New York: Cambridge University Press, 2014), 123–158; John T. Cacioppo, Gary Berntson, John F. Sheridan, et al., "Multilevel Integrative Analyses of Human Behavior: Social Neuroscience and the Complementing Nature of Social and Biological Approaches," *Psychological Bulletin* 126 (6) (2000): 829–843; Tiffany A. Ito and John T. Cacioppo, "Attitudes as Mental and Neural States of Readiness: Using Physiological Measures to Study Implicit Attitudes," in *Implicit Measures of Attitudes*, ed. Bernd Wittenbrink and Norbert Schwarz (New York: The Guilford Press, 2007), 125–158; Nira Liberman, Jens Foerster, and E. Tory Higgins, "Completed vs. Interrupted Priming: Reduced Accessibility from Post-Fulfillment Inhibition," *Journal of Experimental Social Psychology* 43 (2) (2007): 258–264; and Damian A. Stanley and Ralph Adolphs, "Toward a Neural Basis for Social Behavior," *Neuron* 80 (3) (2013): 816–826.

[21] Amodio, "Can Neuroscience Advance Social Psychological Theory?"; Berkman, Cunningham, and Lieberman, "Research Methods in Social and Affective Neuroscience"; and John T. Cacioppo, "Social Neuroscience: Understanding the Pieces Fosters Understanding the Whole and Vice Versa," *American Psychologist* 57 (11) (2002): 819–831.

[22] John T. Cacioppo and Gary G. Berntson, "Social Psychological Contributions to the Decade of the Brain: Doctrine of Multilevel Analysis," *American Psychologist* 47 (8) (1992): 1019–1028; John T. Cacioppo and Curt A. Sandman, "Physiological Differentiation of Sensory and Cognitive Tasks as a Function of Warning, Processing Demands, and Reported Unpleasantness," *Biological Psychology* 6 (3) (1978): 181–192; John T. Cacioppo, "If Attitudes Affect How Stimuli Are Processed, Should They Not Affect the Event-Related Brain Potential?" *Psychological Science* 4 (2) (1993): 108–112; and Monica Fabiani, Gabriele Gratton, and Michael G. H. Coles, "Event-Related Brain Potentials: Methods, Theory," in *Handbook of Psychophysiology*, 2nd edition, ed. John T. Cacioppo, Louis G. Tassinary, and Gary G. Berntson (Cambridge: Cambridge University Press, 2000), 53–84.

[23] Jennifer T. Kubota and Tiffany A. Ito, "You Were Always on My Mind: How Event-Related Potentials Inform Impression Formation Research," in *Handbook of Prejudice, Stereotyping and Discrimination*, ed. T. D. Nelson (New York: Psychology Press, 2009), 279–299; and Bradley D. Mattan, Kevin Y. Wei, Jasmin Cloutier, and Jennifer T. Kubota, "The Social Neuroscience of Race- and Status-Based Prejudice," *Current Opinion in Psychology* 24 (2018): 27–34.

[24] Jennifer T. Kubota and Tiffany Ito, "Rapid Race Perception Despite Individuation and Accuracy Goals," *Social Neuroscience* 12 (4) (2017): 468–478.

[25] James V. Haxby, Elizabeth A. Hoffman, and M. Ida Gobbini, "The Distributed Human Neural System for Face Perception," *Trends in Cognitive Sciences* 4 (6) (2000): 223–233; Nancy Kanwisher, Josh McDermott, and Marvin M. Chun, "The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception," *Journal of Neuroscience* 17 (11) (1997): 4302–4311.

[26] Mattan, Wei, Cloutier, and Kubota, "The Social Neuroscience of Race- and Status-Based Prejudice"; Jennifer T. Kubota, Mahzarin R. Banaji, and Elizabeth A. Phelps, "The Neuroscience of Race," *Nature Neuroscience* 15 (7) (2012): 940–948; Keith B. Senholzi and Jennifer T. Kubota, "The Neural Mechanisms of Prejudice Intervention," in *Neuroimaging Personality, Social Cognition, and Character*, ed. John Absher and Jasmin Cloutier (Amsterdam: Elsevier, 2016), 337–354; Jennifer T. Kubota and Elizabeth A. Phelps, "Insights from Functional Magnetic Resonance Imaging Research on Race," in *Handbook of Prejudice, Stereotyping, and Discrimination*, ed. Todd D. Nelson (Abingdon-on-Thames: Routledge/Psychology Press, 2016), 299–312; and Jennifer T. Kubota and Elizabeth A. Phelps, "Exploring the Brain Dynamics of Racial Stereotyping and Prejudice," in *Social Cognitive Neuroscience, Cognitive Neuroscience, Clinical Brain Mapping* (Amsterdam: Elsevier, 2015), 241–246.

[27] Damian Stanley, Elizabeth A. Phelps, and Mahzarin R. Banaji, "The Neural Basis of Implicit Attitudes," *Current Directions in Psychological Science* 17 (2) (2008): 164–170.

[28] Allen J. Hart, Paul J. Whalen, Lisa M. Shin, et al., "Differential Response in the Human Amygdala to Racial Outgroup Versus Ingroup Face Stimuli," *NeuroReport* 11 (11) (2000): 2351–2354; Matthew D. Lieberman, Ahmad Hariri, Johanna M. Jarcho, et al., "An fMRI Investigation of Race-Related Amygdala Activity in African-American and Caucasian-American Individuals," *Nature Neuroscience* 8 (6) (2005): 720–722; Jaclyn Ronquillo, Thomas F. Denson, Brian Lickel, et al., "The Effects of Skin Tone On Race-Related Amygdala Activity: An fMRI Investigation," *Social Cognitive and Affective Neuroscience* 2 (1) (2007): 39–44; Jennifer A. Richeson, Abigail A. Baird, Heather L. Gordon, et al., "An fMRI Examination of the Impact of Interracial Contact on Executive Function," *Nature Neuroscience* 6 (12) (2003): 1323–1328; and Austen L. Krill and Steven M. Platek, "In-Group and Out-Group Membership Mediates Anterior Cingulate Activation to Social Exclusion," *Frontiers in Evolutionary Neuroscience* 1 (1) (2009).

[29] Elizabeth A. Phelps and Joseph E. LeDoux, "Contributions of the Amygdala to Emotion Processing: From Animal Models to Human Behavior," *Neuron* 48 (2) (2005): 175–187.

[30] Luiz Pessoa and Ralph Adolphs, "Emotion Processing and the Amygdala: From a 'Low Road' to 'Many Roads' of Evaluating Biological Significance," *Nature Reviews Neuroscience* 11 (11) (2010): 1–10; and William A. Cunningham and Tobias Brosch, "Motivational Salience Amygdala Tuning from Traits, Needs, Values, and Goals," *Current Directions in Psychological Science* 21 (1) (2012): 54–59.

[31] Hart, Whalen, Shin, et al., "Differential Response in the Human Amygdala to Racial Outgroup Versus Ingroup Face Stimuli"; Ronquillo, Denson, and Lickel, "The Effects of Skin Tone on Race-Related Amygdala Activity"; Richeson, Baird, Gordon, et al., "An fMRI Examination of the Impact of Interracial Contact on Executive Function"; Krill and Platek, "In-Group and Out-Group Membership Mediates Anterior Cingulate Activation to Social Exclusion"; William A. Cunningham and Tobias Brosch, "Motivational Salience Amygdala Tuning from Traits, Needs, Values, and Goals," *Current Directions in Psychological Science* 21 (1) (2012): 54–59; Jasmin Cloutier, William M. Kelley, and Todd F. Heatherton, "The Influence of Perceptual and Knowledge-Based Familiarity on the Neural Substrates of Face Perception," *Social Neuroscience* 6 (1) (2011): 63–75; Tianyi Li, Carlos Cardenas-Iniguez, Joshua Correll, and Jasmin Cloutier, et al., "The Impact of Motivation on Race-Based Impression Formation," *NeuroImage* 124 (2016): 1–7; Jasmin Cloutier, Tianyi Li, and Joshua Correll, "The Impact of Childhood Experience on Amygdala Response to Perceptually Familiar Black and White Faces," *Journal of Cognitive Neuroscience* 26 (9) (2014): 1992–2004; Richa Gautam, Jasmin Cloutier,

and Jennifer Kubota, "Social Neuroscience of Intergroup Decision-Making," in *Handbook on the Psychology of Morality*, ed. Naomi Ellemers, Stefano Pagliaro, and Félice van Nunspeet (London : Routledge, 2023) ; Jasmin Cloutier, Tianyi Li, Bratislav Miši, et al., "Brain Network Activity During Face Perception : The Impact of Perceptual Familiarity and Individual Differences in Childhood Experience," *Cerebral Cortex* 27 (9) (2016) : 1–13 ; William A. Cunningham, Marcia K. Johnson, Carol L. Raye, et al., "Separable Neural Components in the Processing of Black and White Faces," *Psychological Science* 15 (12) (2004) : 806–813 ; Chad E. Forbes, Christine L. Cox, Toni Schmader, and Lee Ryan, "Negative Stereotype Activation Alters Interaction between Neural Correlates of Arousal, Inhibition and Cognitive Control," *Social Cognitive and Affective Neuroscience* 7 (7) (2012) : 771–781 ; Damian A. Stanley, Peter Sokol-Hessner, Dominic S. Fareri, et al., "Race and Reputation : Perceived Racial Group Trustworthiness Influences the Neural Correlates of Trust Decisions," *Philosophical Transactions of the Royal Society of London B : Biological Sciences* 367 (1589) (2012) : 744–753 ; and Mary E. Wheeler and Susan T. Fiske, "Controlling Racial Prejudice : Social-Cognitive Goals Affect Amygdala and Stereotype Activation," *Psychological Science* 16 (1) (2005) : 56–63.

[32] Elizabeth A. Phelps, Kevin J. O'Connor, William A. Cunningham, et al., "Performance on Indirect Measures of Race Evaluation Predicts Amygdala Activation," *Journal of Cognitive Neuroscience* 12 (5) (2000) : 729–738 ; and Keise Izuma, Ryuta Aoki, Kazuhisa Shibata, and Kiyoshi Nakahara, "Neural Signals in Amygdala Predict Implicit Prejudice toward an Ethnic Outgroup," *Neuroimage* 189 (2019) : 341–352.

[33] Mattan, Wei, Cloutier, and Kubota, "The Social Neuroscience of Race- and Status-Based Prejudice" ; Kubota, Banaji, and Phelps, "The Neuroscience of Race" ; David M. Amodio, "The Neuroscience of Prejudice and Stereotyping," *Nature Reviews Neuroscience* 15 (10) (2014) : 670–682 ; and David M. Amodio and Mina Cikara, "The Social Neuroscience of Prejudice," *Annual Review of Psychology* 72 (2021) : 439–469.

[34] Mattan, Wei, Cloutier, and Kubota, "The Social Neuroscience of Race- and Status-Based Prejudice" ; Li et al., "The Impact of Motivation on Race-Based Impression Formation" ; and Jay J. Van Bavel, Dominic J. Packer, and William A. Cunningham, "The Neural Substrates of In-Group Bias : A Functional Magnetic Resonance Imaging Investigation," *Psychological Science* 19 (11) (2008) : 1131–1139.

[35] Mattan, Wei, Cloutier, and Kubota, "The Social Neuroscience of Race- and Status-Based Prejudice" ; Amodio, "The Neuroscience of Prejudice and Stereotyping" ; Adam M. Chekroud, Jim A. C. Everett, Holly Bridge, and Miles Hewstone, "A Review of Neuroimaging Studies of Race-Related Prejudice : Does Amygdala Response Reflect Threat ?" *Frontiers in Human Neuroscience* 8 (2014) : 179 ; Tanaz Molapour, Armita Golkar, Carlos David Navarrete, et al., "Neural Correlates of Biased Social Fear Learning and Interaction in an Intergroup Context," *NeuroImage* 121 (2015) : 171–183 ; Bradley D. Mattan, Jennifer T. Kubota, Tianyi Li, et al., "Motivation Modulates Brain Networks in Response to Faces Varying in Race and Status : A Multivariate Approach," *eNeuro* 5 (4) (2018) ; and Bradley D. Mattan, Jennifer T. Kubota, Tzipporah P. Dang, and Jasmin Cloutier, "External Motivation to Avoid Prejudice Alters Neural Responses to Targets Varying in Race and Status," *Social Cognitive and Affective Neuroscience* 13 (1) (2017) : 22–31.

[36] Cloutier, Li, and Correll, "The Impact of Childhood Experience on Amygdala Response to Perceptually Familiar Black and White Faces."

[37] Ito and Kubota, "The Social Neuroscience of Social Cognition" ; Kubota and Ito, "You Were Always on My Mind" ; and David M. Amodio, Jennifer T. Kubota, Eddie Harmon-Jones, and Patricia G. Devine, "Alternative Mechanisms for Regulating Racial Responses

According to Internal vs. External Cues," *Social Cognitive and Affective Neuroscience* 1 (1) (2006): 26–36.

[38] Matthew M. Botvinick, Todd S. Braver, Deanna M. Barch, et al., "Conflict Monitoring and Cognitive Control," *Psychological Review* 108 (3) (2001): 624–652.

[39] Amitai Shenhav, Jonathan D. Cohen, and Matthew M. Botvinick, "Dorsal Anterior Cingulate Cortex and the Value of Control," *Nature Neuroscience* 19 (10) (2016): 1286–1291.

[40] Brittany S. Cassidy, Gregory T. Sprout, Jonathan B. Freeman, and Anne C. Krendl, "Looking the Part (to Me): Effects of Racial Prototypicality on Race Perception Vary by Prejudice," *Social Cognitive and Affective Neuroscience* 12 (4) (2017): 685–694.

[41] Richeson, Baird, Gordon, et al., "An fMRI Examination of the Impact of Interracial Contact on Executive Function"; Cunningham, Johnson, Raye, et al., "Separable Neural Components in the Processing of Black and White Faces"; Jennifer S. Beer, Mirre Stallen, Michael V. Lombardo, et al., "The Quadruple Process Model Approach to Examining the Neural Underpinnings of Prejudice," *NeuroImage* 43 (4) (2008): 775–783; Joseph E. Dunsmoor, Jennifer T. Kubota, Jian Li, et al., "Racial Stereotypes Impair Flexibility of Emotional Learning," *Social Cognitive and Affective Neuroscience* 11 (9) (2016): 1363–1373; and Melike M. Fourie, Kevin G. F. Thomas, David M. Amodio, et al., "Neural Correlates of Experienced Moral Emotion: An fMRI Investigation of Emotion in Response to Prejudice Feedback," *Social Neuroscience* 9 (2) (2014): 203–218.

[42] Amodio, Kubota, Harmon-Jones, and Devine, "Alternative Mechanisms for Regulating Racial Responses According to Internal vs. External Cues"; David M. Amodio, James Y. Shah, Jonathan Sigelman, et al., "Implicit Regulatory Focus Associated with Asymmetrical Frontal Cortical Activity," *Journal of Experimental Social Psychology* 40 (2) (2004): 225–232; and David M. Amodio, Patricia G. Devine, and Eddie Harmon-Jones, "Individual Differences in the Regulation of Intergroup Bias: The Role of Conflict Monitoring and Neural Signals for Control," *Journal of Personality and Social Psychology* 94 (1) (2008): 60–74.

[43] Richeson, Baird, Gordon, et al., "An fMRI Examination of the Impact of Interracial Contact on Executive Function"; and Brittany S. Cassidy and Anne C. Krendl, "Dynamic Neural Mechanisms Underlie Race Disparities in Social Cognition," *NeuroImage* 132 (2016): 238–246.

[44] Kartik K. Sreenivasan, Clayton E. Curtis, and Mark D'Esposito, "Revisiting the Role of Persistent Neural Activity during Working Memory," *Trends in Cognitive Sciences* 18 (2) (2014): 82–89.

[45] Brittany S. Cassidy, Eunice J. Lee, and Anne C. Krendl, "Age and Executive Ability Impact the Neural Correlates of Race Perception," *Social, Cognitive, and Affective Neuroscience* 11 (11) (2016): 1752–1761.

[46] Michael W. L. Chee, Natarajan Sriram, Chun Siong Soon, and Kok Ming Lee, "Dorsolateral Prefrontal Cortex and the Implicit Association of Concepts and Attributes," *NeuroReport* 11 (1) (2000): 135–140.

[47] Mattan, Wei, Cloutier, and Kubota, "The Social Neuroscience of Race- and Status-Based Prejudice"; Kubota, Banaji, and Phelps, "The Neuroscience of Race"; Amodio, "The Neuroscience of Prejudice and Stereotyping"; and Amodio and Cikara, "The Social Neuroscience of Prejudice."

[48] Cloutier, Kelley, and Heatherton, "The Influence of Perceptual and Knowledge-Based Familiarity on the Neural Substrates of Face Perception."

[49] Elizabeth A. Phelps, Kevin J. O'Connor, William A. Cunningham, et al., "Performance on Indirect Measures of Race Evaluation Predicts Amygdala Activation," *Journal of Cognitive Neuroscience* 12 (5) (2000): 729–738.

[50] Cloutier, Li, and Correll, "The Impact of Childhood Experience on Amygdala Response to Perceptually Familiar Black and White Faces."

[51] Eva H. Telzer, Kathryn L. Humphreys, Mor Shapiro, and Nim Tottenham, "Amygdala Sensitivity to Race Is Not Present in Childhood but Emerges over Adolescence," *Journal of Cognitive Neuroscience* 25 (2) (2012): 234–244.

[52] Reginald B. Adams, Nicholas O. Rule, Robert G. Franklin Jr., et al., "Cross-Cultural Reading the Mind in the Eyes: An fMRI Investigation," *Journal of Cognitive Neuroscience* 22 (1) (2010): 97–108; David M. Amodio and Chris D. Frith, "Meeting of Minds: The Medial Frontal Cortex and Social Cognition," *Nature Reviews Neuroscience* 7 (4) (2006): 268–277; Jason P. Mitchell, C. Neil Macrae, and Mahzarin R. Banaji, "Dissociable Medial Prefrontal Contributions to Judgments of Similar and Dissimilar Others," *Neuron* 50 (4) (2006): 655–663; Rebecca Saxe and Lindsey J. Powell, "It's the Thought That Counts: Specific Brain Regions for One Component of Theory of Mind," *Psychological Science* 17 (8) (2006): 692–699; Frank Van Overwalle, "Social Cognition and the Brain: A Meta-Analysis," *Human Brain Mapping* 30 (3) (2009): 829–858; and Rebecca Saxe, "Why and How to Study Theory of Mind with fMRI," *Brain Research* 1079 (1) (2006): 20.

[53] Grace Handley, Jennifer T. Kubota, and Jasmin Cloutier, "Reading the Mind in the Eyes of Black and White People: Interracial Contact and Perceived Race Affects Brain Activity When Inferring Mental States," *NeuroImage* 269 (2023): 119910; and Grace Handley, Jennifer Kubota, and Jasmin Cloutier, "Interracial Contact Differentially Shapes Brain Networks Involved in Social and Non-Social Judgments from Faces: A Combination of Univariate and Multivariate Approaches," *Social Cognitive and Affective Neuroscience* 17 (2) (2022): 218–230.

[54] Tzipporah P. Dang, Bradley D. Mattan, Denise M. Barth, et al., "Perceiving Social Injustice during Arrests of Black and White Civilians by White Police Officers: An fMRI Investigation," *NeuroImage* 255 (2022): 119153; and Jennifer T. Kubota, Tzipporah P. Dang, Bradley D. Mattan, et al., "Social Justice Neuroscience, a Valuable and Complex Endeavor: Authors' Reply to Commentaries on 'Perceiving Social Injustice during Arrests of Black and White Civilians by White Police Officers: An fMRI Investigation,'" *NeuroImage* 255 (2022): 119155.

[55] Handley, Kubota, and Cloutier, "Interracial Contact Differentially Shapes Brain Networks Involved in Social and Non-Social Judgments from Faces."

[56] Bertram Gawronski, "Six Lessons for a Cogent Science of Implicit Bias and Its Criticism," *Perspectives on Psychological Science* 14 (4) (2021): 574–595.

[57] Calvin K. Lai, Maddalena Marini, Steven A. Lehr, et al., "Reducing Implicit Racial Preferences: I. A Comparative Investigation of 17 Interventions," *Journal of Experimental Psychology: General* 143 (4) (2014): 1765–1785; Patrick S. Forscher, Calvin K. Lai, Jordan R. Axt, et al., "A Meta-Analysis of Changes in Implicit Bias" (unpublished manuscript, 2016); Evelyn R. Carter, Ivuoma N. Onyeador, and Neil A. Lewis Jr., "Developing & Delivering Effective Anti-Bias Training: Challenges & Recommendations," *Behavioral Science & Policy* 6 (1) (2020): 57–70; and Ivuoma N. Onyeador, Sa-kiera T. J. Hudson, and Neil A. Lewis, Jr., "Moving beyond Implicit Bias Training: Policy Insights for Increasing Organizational Diversity," *Policy Insights from the Behavioral and Brain Sciences* 8 (1) (2021): 19–26.

[58] Malte Friese, Wilhelm Hofmann, and Manfred Schmitt, "When and Why Do Implicit Measures Predict Behaviour? Empirical Evidence for the Moderating Role of Opportunity, Motivation, and Process Reliance," *European Review of Social Psychology* 19 (1) (2008): 285–338; Marco Perugini, Juliette Richetin, and Cristina Zogmaister, "Prediction of Behavior," in *Handbook of Implicit Social Cognition*, ed. Gawronski and Payne; Kurdi, Seitchik, Axt, et al., "Relationship between the Implicit Association Test and Intergroup Behavior"; Greenwald, Poehlman, Uhlmann, and Banaji, "Understanding and Using the Implicit Association Test"; C. Daryl Cameron, Jazmin L. Brown-Iannuzzi, and B. Keith Payne, "Sequential Priming Measures of Implicit Social Cognition: A Meta-Analysis of Associations with Behavior and Explicit Attitudes," *Personality and Social Psychology Review* 16 (4) (2012): 330–350; and Frederick L. Oswald, Gregory Mitchell, Hart Blanton, et al., "Using the IAT to Predict Ethnic and Racial Discrimination: Small Effect Sizes of Unknown Societal Significance," *Journal of Personality and Social Psychology* 108 (4) (2015): 562–571.

[59] Hart Blanton and James Jaccard, "You Can't Assess the Forest If You Can't Assess the Trees: Psychometric Challenges to Measuring Implicit Bias in Crowds," *Psychological Inquiry* 28 (4) (2017): 249–257; and Gregory Mitchell, "An Implicit Bias Primer," *Virginia Journal of Social Policy & the Law* 25 (1) (2018): 27–58.

[60] Brian A. Nosek, Carlee Beth Hawkins, and Rebecca S. Frazier, "Implicit Social Cognition: From Measures to Mechanisms," *Trends in Cognitive Sciences* 15 (4) (2011): 152–159.

[61] Mahzarin R. Banaji, Susan T. Fiske, and Douglas S. Massey, "Systemic Racism: Individuals and Interactions, Institutions and Society," *Cognitive Research: Principles and Implications* 6 (2021): 1–21; and Mark L. Hatzenbuehler et al., "Community-Level Explicit Racial Prejudice Potentiates Whites' Neural Responses to Black Faces: A Spatial Meta-Analysis," *Social Neuroscience* 17 (6) (2022): 1–12.

[62] Bobo, "Racial Attitudes and Relations at the Close of the Twentieth Century"; Banaji, Fiske, and Massey, "Systemic Racism: Individuals and Interactions, Institutions and Society"; and Eduardo Bonilla-Silva and Amanda Lewis, *The "New Racism": Toward an Analysis of the U.S. Racial Structure, 1960s – 1990s* (Ann Arbor: Center for Research on Social Organization at the University of Michigan, 1996).

[63] Kubota, Dang, Mattan, et al., "Social Justice Neuroscience, a Valuable and Complex Endeavor."

[64] Richa Gautam, Jasmin Cloutier, and Jennifer Kubota, "Social Neuroscience of Intergroup Decision-Making," in *Handbook on the Psychology of Morality*, ed. Naomi Ellemers, Stefano Pagliaro, and Félice van Nunspeet (London: Routledge, 2023).

[65] Kubota, Banaji, and Phelps, "The Neuroscience of Race."

[66] Sylvia Terbeck, Guy Kahane, Sarah McTavish, et al., "Propranolol Reduces Implicit Negative Racial Bias," *Psychopharmacology* 222 (3) (2012): 419–424.

[67] Raamy Majeed, "On Biologizing Racism," *The British Journal for the Philosophy of Science* (forthcoming).

Implicit Bias as a Cognitive Manifestation of Systemic Racism

Manuel J. Galvan & B. Keith Payne

Explicitly prejudiced attitudes against Black Americans have declined gradually since the 1960s. Yet racial disparities and racial discrimination remain significant problems in the United States. How could discrimination and disparate outcomes remain constant even while racial prejudice decreased? Two prominent explanations have emerged to explain these puzzling trends. Sociologists have proposed that disparities and discrimination are perpetuated by systemic racism, or the policies, practices, and societal structures that disadvantage some racial groups compared with others. Simultaneously, psychologists have proposed that implicit biases may sustain discrimination even in the absence of explicit prejudice. In this essay, we explore newly discovered connections between systemic racism and implicit bias, how they challenge traditional views to reorient our understanding of implicit bias, and how they shed new light on strategies to reduce bias.

In 2022, artificial intelligence researchers at OpenAI released their latest development, ChatGPT. Using machine learning algorithms trained on large bodies of text, the chatbot could generate impressively human-sounding text responses on seemingly endless topics. Users soon began debating whether the technology had reached human-like levels of intelligence, even going so far as to invoke the concept of sentience.[1] Meanwhile, those with experience using artificial intelligence worried about a problem that has plagued the field for years: chatbots trained on human inputs are prone to saying racist, sexist, and otherwise offensive things.

The designers of ChatGPT had anticipated this problem with bias and had installed new filters to prevent the bot from saying inappropriate things. If you asked ChatGPT to tell a racist joke, for example, the bot would refuse, explaining: "I am not capable of generating offensive or harmful content." But as cognitive scientist Steven Piantadosi noted, those biased inferences were still there, and could be revealed by probing indirectly.[2] When he asked the chatbot to write a computer code function to check if someone would be a good scientist based on their race and gender, it generated code indicating that only white males were good scientists. When asked to create code to decide if a child's life should be saved based on their race and gender, the code indicated that the lives of Black males should not be saved.

https://doi.org/10.1162/daed_a_02051

As psychologists, we find that the continuing struggle to create artificial intelligence that is free from racism says more about humanity than about technology. The algorithms behind this chatbot make statistical predictions about what words go together, based on training with massive bodies of real-world text. When a statistical model returns a biased response, it reflects the biases in the human texts on which it was trained. Programmers can add rules like "don't say racist things," but that does not change the biases that are deeply embedded in the training environment. As a result, the chatbot may seem unbiased when asked directly but will reveal biases indirectly in countless ways. Artificial intelligence has thus encapsulated what psychologists have known about humans for decades: when a cognitive system that detects statistical regularities is immersed in an environment that is systemically biased, it will reproduce those biases.

The chatbot highlights something else about human psychology. When a robot reproduces biases, it is easy for humans to identify its environment as the source of the bias. Few people believe that there is something deep and essential about the robot's character that makes it racist. When humans form the same kinds of biased associations, however, people tend to attribute it to the attitudes, beliefs, or character of the person.

We argue that the human mind, like artificial intelligence, tracks statistical regularities in the social environment. When the mind is immersed in an environment of systemic racism, it tends to form biased associations and inferences about marginalized social groups. In fact, implicit bias is best understood as the cognitive reflection of systemic racism. This formulation may seem surprising: implicit bias has long been thought of as an individual trait or attitude, whereas systemic racism concerns structures, history, and social environments, rather than individuals. In this essay, we explore the connection between systemic racism and implicit bias: how it challenges traditional views to reorient our understanding of implicit bias, and how it sheds new light on strategies to reduce bias.

The theory of implicit bias grew out of efforts to understand gradual trends toward more egalitarian attitudes in standardized surveys. For example, beginning in the 1960s, white Americans have slowly caught up with Black Americans on issues of interpersonal discrimination. Today, over 90 percent of white and Black Americans support racially integrated schooling and reject laws against interracial marriage.[3] Another poll in 2019 found that 72 percent of white respondents believe it is never acceptable for a white person to use the N-word.[4] Such polling data illustrate the eventual decline in white people expressing explicit biases against Black Americans in surveys.

At the same time, actual racial disparities have remained largely undiminished. Relative to white Americans, Black Americans are far more likely to struggle with poverty, food insecurity, and unemployment.[5] Black Americans have 10 percent of

the median net worth and half the median annual income of their white counterparts.[6] Such disparities are hard to address when racial discrimination persists. For decades, researchers have conducted field experiments responding to job postings with two versions of otherwise identical résumés: one with a name that implies a Black identity and the other with a name that implies a white identity. The rate of callbacks to the applicants is a measure of racial discrimination between otherwise equally qualified candidates. Recent meta-analyses of similar field experiments have demonstrated that racial discrimination in hiring has remained relatively constant since the late 1980s, and housing discrimination has decreased but remains potent.[7]

These trends created a puzzle. How could discrimination and disparate outcomes remain constant even while racial prejudice decreased? This question spurred innovations in thinking across the social sciences.

Sociologists developed the concept of systemic racism to account for the ways that inequalities can be perpetuated independent of individuals' attitudes and intentions. *Systemic racism* refers to policies, practices, and societal structures that disadvantage some racial groups compared with others.[8] This is distinct from more colloquial uses of the word racism to describe prejudicial thoughts, beliefs, or behaviors, which is often referred to as *interpersonal racism*. An essential theoretical contribution of systemic racism research is the recognition that individual actors do not simply act as racists or nonracists. For example, even if all discriminatory behavior stopped today, preexisting disparities in income, wealth, and educational opportunity would still ensure that racial inequalities are passed on to future generations.

Psychologists grappled with the puzzle of persistent discrimination and disparities amid attitudinal shifts toward egalitarianism by developing the concept of implicit bias. Implicit bias refers to positive or negative mental associations cued spontaneously by social groups. It is measured using cognitive tasks that test how those associations facilitate or interfere with task performance. Unlike survey methods, implicit tests are difficult to manipulate based on social desirability or norms against expressing prejudice. Studies suggest that implicit bias is widespread, even among people who explicitly endorse egalitarian attitudes.[9] If implicit bias leads to discriminatory behavior, it could explain the puzzle of widespread discrimination despite declining prejudice on surveys.

Implicit bias has traditionally been considered an individual attitude. Implicit bias tests and sequential priming tasks were developed as individual difference measures.[10] Most theories of implicit bias posit that implicit attitudes were learned from cultural biases early in development and became rigid because of immense repetition.[11]

The ideas of systemic racism and implicit bias were thus developed as very different approaches to solving the same puzzle. One focused on the ways that laws,

policies, and social environments perpetuate inequalities without regard to individual attitudes. The other focused almost entirely on individual attitudes. However, recent research has reconsidered implicit bias as an individual trait. We argue that these two theoretical frameworks are not as different as was once assumed, and that implicit bias is in fact a cognitive reflection of systemic racism in the environment.

As research accumulated over the past two decades, several findings cast doubt on the individual-attitude view of implicit bias. For one, we expect individual attitudes to be stable over time. Implicit biases can be reliably detected in group averages using tests such as the Implicit Association Test (IAT) or the Affect Misattribution Procedure (AMP).[12] If one hundred randomly selected Americans completed the IAT, there is a very high likelihood that there would be a detectable average level of implicit bias across the group. However, longitudinal studies have found that while group averages are consistent over time, individual scores are quite unstable.[13] In other words, when a classroom of students takes the IAT, the rank order of the students will change such that the most and least biased students may not be the same when they retake the IAT at the end of the semester. Yet the classroom average will remain remarkably similar.

A second related puzzle is that average implicit bias does not change over the lifespan. Groups of younger and older Americans of various ages have been found to have very similar average implicit biases.[14] Under the traditional attitude assumption, this stability would naturally result from stable individual biases. However, given the temporal instability of individual-level bias, this age invariance is surprising. How can a variable that is unstable over two weeks be stable across a lifetime?

A third puzzling finding is that implicit biases of individuals are not strong predictors of individual discriminatory behavior (r = 0.14 to 0.24).[15] Yet when implicit biases are aggregated over larger geographic areas, they have much stronger associations to behavioral outcomes such as achievement gaps, disparities in shootings, health disparities, and internet searches using racial slurs.[16]

In light of these anomalies, psychologists B. Keith Payne, Heidi A. Vuletich, and Kristjen B. Lundberg developed the "bias of crowds" model to make sense of the large body of implicit bias research.[17] The basic assumption of the model is that implicit bias scores reflect the accessibility of concepts linked to social group categories. Concept associations can vary both chronically, as an individual difference from one person to another, and situationally, from one context to the next. For very stable individual constructs, like explicit racial attitudes, there is a lot of stable individual variation but little temporal variation within persons. Some people have explicit biases, others don't, but each individual's explicit biases are generally consistent over time. When stable traits are aggregated, the aggre-

gate measure's stability is simply a reflection of the stability in individual differences in scores.

Despite its capriciousness at the individual level, implicit bias can be remarkably stable at the context level (such as city, county, or state level). We describe this as *emergent stability* because the aggregate stability cannot be reduced to stability of the individual scores. The ranking of people from highest to lowest implicit biases will shuffle over time, yet there will be a consistent mean level of implicit bias for the group. The bias of crowds model suggests that this consistent emergent stability reflects the relatively stable social context. Features of context make implicit associations between racial groups more or less prominent. When aggregation occurs, random variation at the individual level is reduced, enabling a clearer estimation of the influence of shared contextual factors on implicit bias. Given its emergent stability, implicit bias at the context level becomes a more theoretically and practically useful predictor and outcome for social scientists.

Many variables can be either measured as individual differences or averaged across individuals to measure contexts. For example, very stable aspects of personality have been found to vary across geographic regions. People in Middle America and the South (typically "red states") are more inclined to be "friendly and conventional," meaning they are higher in conscientiousness, agreeableness, and extraversion, and low in neuroticism and openness.[18] Regularities in regional personality structure are thought to be due to regularities in the physical environment, historical events, and cultural norms of the region.[19]

But unlike very stable personality constructs, implicit bias is limited as an individual difference variable, and is instead particularly powerful as a context-based measure. One reason, reviewed above, is that implicit bias scores are very low in stability. A second related reason is that implicit bias scores are highly context sensitive.[20] For example, experiments have shown that seeing Black Americans in a positively valenced context, like at church or a family barbecue, results in participants having lower anti-Black implicit biases compared with when they see Black Americans in a negative context, like prison.[21]

Because implicit bias is unstable and highly context sensitive, the average implicit bias in a city, county, or state is not reducible to the attitudes of the individuals that make up the context. This means that when we take a sample of participants from a given context to measure their implicit bias, the specific individuals in our sample are largely interchangeable. If you replaced the individuals sampled with another set of individuals from the same context, their aggregated scores would show the same level of implicit bias. Whatever influence is exerted by the context will be reflected in the scores of whoever inhabits those spaces. Because of this, aggregated implicit bias scores have proved to be extraordinarily sensitive indicators of systemic racism.

Much research suggests that implicit biases are influenced by contextual information in the environment. There is a large body of literature showing that implicit associations are influenced by experimental procedures. For example, one study attempted to influence the association between Middle Easterners and negative words by exposing participants to a slideshow showing Middle Eastern faces paired with positive images and white faces paired with neutral images. Relative to a control group shown the same images but without pairing the stimuli, the experimental participants showed a lessened degree of Middle Eastern implicit bias.[22] A meta-analysis of more than two hundred studies performed over many decades showed such evaluative conditioning effects on implicit biases are replicable, if small.[23] Other studies have demonstrated that counter-stereotypical experiences, such as positive interactions with a Black experimenter or reading about positive exemplars, can reduce negative implicit biases.[24] These laboratory studies demonstrate that implicit biases are subject to significant shifts due to environmental conditions. Conditions that match cultural stereotypes of marginalized groups strengthen associations with negative concepts and reinforce implicit biases.

Outside of the laboratory, systemic racism, as a set of long-standing structural, institutional, and cultural tendencies, has created the specific environmental conditions that would theoretically reinforce implicit biases. Most Black Americans are descendants of enslaved African people brought to the continent prior to the abolition of slavery.[25] The slave trade was a four-centuries-long brutal and dehumanizing regime that included capture, enslavement, destruction of African identity, disruption of families, and indoctrination of Black inferiority. Such trauma was also perpetuated by intergenerational familial trauma.[26] The legacy of slavery can be seen in contemporary patterns of distrust between ethnic groups, voting behavior, and cultural norms, belief, and values.[27] It also set the stage for the enormous wealth gap between white and Black people in the United States that has not meaningfully closed.[28]

While the Thirteenth Amendment officiated the end of slavery by federal law, there is a complex and sordid history between abolition in 1865 and the civil rights era in the late 1960s. In that time, systemic racism was a brazen and institutionalized set of practices that included Black Codes, sharecropping, lynching, Jim Crow laws, sundown towns, and redlining. Many studies have tried to estimate how these structures and events have shaped the contemporary context of Black-white inequality.

An analysis of U.S.-based health outcomes found that Jim Crow laws had an enduring impact on Black-white mortality rates from 1960 to 2009.[29] Southern counties with higher rates of historical lynchings from 1882 to 1930 had lower Black voter registration in modern elections.[30] Spatial proximity to sundown towns (that is, towns where Black people were subject to violence if they were present after sundown) predicts Black-white poverty disparities.[31] A number of

studies have connected redlining, a legal practice until the passing of the Fair Housing Act in 1968, to current inequality. To provide only a sampling of recent research, historical redlining patterns are associated with life expectancy, the proportion of health care professionals, access to quality food, home heat vulnerability, environmental racism, cardiometabolic risk, tobacco retailer density, gentrification, alcohol outlet density, nonfatal shooting incidence, air pollution, fatal encounters with police, and COVID-19 exposure.[32]

These studies present strong empirical evidence that systemic racism has shaped the life outcomes of both Black and white populations in the United States. In other words, systemic racism is an important contextual factor that strongly influences who is successful, who has educational opportunities, who is exposed to violence and addiction, who lives in expensive homes and communities, and who languishes in poverty and within the carceral system. Such statistical regularities in our society are readily perceived as we walk to work, watch the news, or drive through segregated neighborhoods. For those of us in racially unequal regions of the country, which have been most impacted by systemic racism, there are myriad constant cues that one group has what the other group does not. The bias of crowds model suggests that the context of persistent and systematized inequality between racialized groups underlies the implicit associations we make between racialized groups and concepts like "good," "bad," "criminal," "smart," and "dumb" as measured by instruments like the IAT.

If implicit bias is an indicator of systemic racism, we would expect to find reliable associations between contextual aspects of systemic racism and implicit bias. Some studies consider which aspects of historical and current context might predict higher implicit bias in different geographic regions. As discussed previously, slavery has profoundly influenced current-day culture, behavior, wealth distribution, and other aspects of systemic racism; we would expect that it also underlies implicit biases. This is exactly what research in our lab has found: the historical proportion of enslaved populations at the county and state level predicts implicit bias today.[33] Places that relied on Black slave labor before abolition exhibit today higher pro-white bias among the white residents and lower pro-white bias among the Black residents. This effect persists even after controlling for self-reported attitudes. As we would predict from the bias of crowds theory, the relationship between the proportion of enslaved populations and implicit bias was mediated by structural inequalities like the proportion of Black people and white people in poverty, residential segregation, and intergenerational mobility of Black people and white people. Slavery and the ensuing generations of racial segregation and economic deprivation build the statistical regularities of racial inequality into the context. Chronic exposure to these structural inequalities maintains and exacerbates implicit bias.

Unfortunately, implicit biases are not merely cognitive reflections of our environment. Rather, they are influential aspects of our cognitive processes that

change our behavior. The bias of crowds model suggests a recursive process such that inequalities of the past create the conditions for implicit biases to develop; and when they do, implicit biases contribute to the perpetuation of inequalities going forward. Said another way, implicit bias may be understood as both a cause and an effect of racial inequality.

There are many studies demonstrating that regional differences in implicit bias are associated with an increase in behaviors and outcomes that reinforce racial disparities. Such effects begin before children are born. An analysis of data from thirty-one million births across the United States found that the white-Black disparity in low birth weight is 14 percent higher in counties with high implicit bias.[34] During the SARS-CoV-2 global pandemic, anti-Black implicit biases of the white population across 957 counties predicted higher white and Black incidence of COVID-19 infection and a larger Black-white infection rate gap.[35] These are just specific instances of the larger pattern of racial health disparities following from geographic differences in implicit bias. A systematic review of the literature found evidence that all-cause mortality, cause-specific mortality, birth outcomes, cardiovascular outcomes, mental health, and self-rated health of racially minoritized groups are adversely affected by implicit biases.[36]

Implicit biases are also associated with the experiences of children. Counties with high levels of implicit bias show higher Black-white disparities in disciplinary suspension rates, and counties with higher levels of implicit bias among educators showed higher Black-white disparities in test scores and suspensions (after adjusting for several county-level covariates).[37] Similarly, county-level rates of anti-Black bias predict Black-white disparities in K–12 enrollment in gifted and talented programs such that high levels of bias predict large gaps and low levels of bias predict no gap.[38] U.S. states with higher levels of anti-Black implicit bias are also more likely to have lower adoption rates of Black foster children.[39]

Finally, several studies have demonstrated that regional implicit bias influences policing policy and behavior. Counties with higher anti-Black implicit bias have greater racial disparities in traffic stops.[40] Data from over two million residents across the United States also found that implicit biases predict more police officer use of lethal force against Black residents relative to the base rate of Black people in the population.[41] Researchers have also linked implicit bias to the problem of police militarization: regional differences in prejudice (including implicit bias) predict greater tax allocations for purchasing militarized police equipment.[42]

More research is needed to disentangle the many related factors involved in explaining racial disparities. Much of this work is relatively new and still developing. There is also some apparent overlap between the research linking historical events and policies to implicit biases and structural inequality. As we have recently argued, future researchers need to consider novel ways of incorporating these different factors into a coherent theoretical and statistical model.[43] Doing so will

require collaboration between scholars from different fields like sociology, history, and policy in order to better account for the role of historical events, structural inequalities, and policy regimes, in addition to the usual set of predictors (implicit bias, explicit bias, and demographics) and policing, educational, and economic outcomes.

To address the massive public policy problems of racism and racial inequality in policing, education, economics, and health, racial justice advocates have turned to implicit bias as a focal point for intervention. Generally, this has taken the form of implicit bias trainings, whereby participants engaged in activities – such as perspective-taking, considering counter-stereotypical exemplars, meditating, or viewing empathy-building stimuli – designed to reduce implicit biases.[44] Unfortunately, meta-analyses and large-scale replications of such interventions have demonstrated that while they can successfully reduce implicit biases in the immediate time after intervention, they rarely have a sustained effect on implicit bias.[45] From a bias of crowds perspective, it is unsurprising that such interventions do not have lasting effects. If implicit bias is an emergent property of racial inequality in the social context, interventions that do not change the social context should leave implicit biases relatively unchanged.

By recognizing that context shapes implicit bias and behavior, researchers, policymakers, and practitioners can consider changing the context to reduce implicit bias.[46] At the highest level are societal-scale interventions that would redress historical and current inequality and thus radically change the context. Economist Ellora Derenoncourt and colleagues used economic data from 1860 to 2020 to simulate how economic conditions and policies influence the Black-white wealth gap.[47] Their analyses reveal that different combinations of policies that increase stock (such as lump sum reparations) and flow (such as facilitating financial diversification, stock equity, financial literacy, saving behaviors, and improving educational and labor market outcomes) in the Black community are feasible mechanisms for reducing wealth inequality over the coming decades.

On a smaller scale, individual organizations can reduce implicit bias by shifting organizational policies. Rather than having counter-stereotypical examples embedded in implicit bias training materials, organizations can work to have more counter-stereotypical minoritized group members in their ranks. Having an inclusive, equitable, and diverse team is a way to counteract the maintenance of negative intergroup biases.[48] This approach requires that organizations contend with biases in the hiring process that may hinder the hiring potential of racialized minority group members. To reduce the influence of implicit biases, decision-making processes can be predetermined and specified using hiring rubrics.[49] Though such rubrics can improve the hiring process, they need to follow evidence-based implementation to avoid perpetuating bias.[50] Another approach is to

build in monitoring and accountability in hiring practices. Decision-makers who are held accountable for evaluating job candidates tend to show less pro-white biases.[51] Finally, people are more likely to be influenced by implicit biases when they are rushed, tired, distracted, or over-worked.[52] In a study of more than one thousand three hundred field experiments in classrooms, researchers found that discrimination rates against students from ethnic minority backgrounds were much lower when teachers were provided more time and resources in the classroom.[53]

Shifts in governmental, social, and workplace policies may be more challenging to implement compared with providing an implicit bias training, but policy changes may address the roots of the problem in ways that simple trainings cannot. Generations of public and organizational policy decisions resulted in the racial inequality we have today; the evidence suggests that we need equitable policies to counteract those effects.

Historically, the study of racism in psychological research has largely focused on interpersonal racism and has generally construed racism as an aspect of individual psychology, while neglecting the historical and structural aspects of racism.[54] The bias of crowds model is a theoretical framework that explains why the modern shift toward racial egalitarianism in attitudes has not resulted in diminished racial inequality. It also accounts for the many research findings that are inconsistent with the perspective that implicit bias is a stable aspect of individual psychology.

The other benefit of the bias of crowds model is that it makes efficient use of existing data and theory. Research that links together policy, structural inequality, and psychological measures has been limited by the availability of geo-coded "big data" on these topics. The recent explosion of research linking widescale policies like redlining to health outcomes (for example) is in no small part due to the increased availability of such data. Analyses using these data reveal more evidence that the bias of crowds model is a social psychological model consistent with the sociological theory of systemic racism. Ultimately, the model connects many forms of racism – structural, systemic, implicit, explicit, cultural, historical, current – under one testable theoretical perspective.

Finally, the bias of crowds model reinforces what many in sociology, economics, history, and policy have articulated in their work: we need to consider systems to understand and ameliorate racial inequality. The bias of crowds model shifts the focus of research designed to address inequality to consider the impact of changing the broader social context.

ABOUT THE AUTHORS

Manuel J. Galvan is a PhD candidate in social psychology at the University of North Carolina at Chapel Hill. He is the Cofounder and Cohost of the *A Bit More Complicated* science podcast. He has published in such journals as *Science*, *Annals of Behavioral Medicine*, and *American Psychologist*.

B. Keith Payne is Professor of Psychology and Neuroscience at the University of North Carolina at Chapel Hill. He is the author of *The Broken Ladder: How Inequality Affects the Way We Think, Live, and Die* (2017) and has recently published in such journals as *Psychological Science*, *Journal of Experimental Psychology*, and *Proceedings of the National Academy of Sciences*.

ENDNOTES

[1] Eric Holloway, "Yes, ChatGPT Is Sentient – Because It's Really Humans in the Loop," *Mind Matters*, December 26, 2022, https://mindmatters.ai/2022/12/yes-chatgpt-is-sentient-because-its-really-humans-in-the-loop.

[2] Tony H. Tran, "OpenAI's Impressive New Chatbot Isn't Immune to Racism," *The Daily Beast*, December 6, 2022, https://www.thedailybeast.com/openais-impressive-chatgpt-chatbot-is-not-immune-to-racism.

[3] Sarah Patton Moberg, Maria Krysan, and Deanna Christianson, "Racial Attitudes in America," *Public Opinion Quarterly* 83 (2) (2019): 450–471, https://doi.org/10.1093/poq/nfz014.

[4] Juliana M. Horowitz, Anna Brown, and Kiana Cox, "Race in America 2019," Pew Research Center, April 9, 2019, https://www.pewsocialtrends.org/wp-content/uploads/sites/3/2019/04/PewResearchCenter_RaceStudy_FINAL-1.pdf.

[5] Economic Research Service, *Food Security Status of U.S. Households in 2021* (Washington, D.C.: Economic Research Service, 2023), https://www.ers.usda.gov/topics/food-nutrition-assistance/food-security-in-the-u-s/key-statistics-graphics/#householdtype; Jonnelle Marte, "Gap in U.S. Black and White Unemployment Rates Is Widest in Five Years," Reuters, July 2, 2020, https://www.reuters.com/article/us-usa-economy-unemployment-race/gap-in-u-s-black-and-white-unemployment-rates-is-widest-in-five-years-idUSKBN2431X7; and U.S. Census Bureau, "Historical Poverty Tables: People and Families – 1959 to 2021," last modified January 30, 2023, https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-poverty-people.html.

[6] Lisa Dettling, Joanne Hsu, Lindsay Jacobs, Kevin B. Moore, and Jeffrey Thompson, "Recent Trends in Wealth-Holding by Race and Ethnicity: Evidence from the Survey of Consumer Finances," FEDS Notes, September 27, 2017, https://www.federalreserve.gov/econres/notes/feds-notes/recent-trends-in-wealth-holding-by-race-and-ethnicity-evidence-from-the-survey-of-consumer-finances-20170927.html; and Valerie Wilson and Jhacova Williams, "Racial and Ethnic Income Gaps Persist Amid Uneven Growth in Household Incomes," Economic Policy Institute Working Economics Blog, September 11, 2019, https://www.epi.org/blog/racial-and-ethnic-income-gaps-persist-amid-uneven-growth-in-household-incomes.

[7] Katrin Auspurg, Andreas Schneck, and Thomas Hinz, "Closed Doors Everywhere? A Meta-Analysis of Field Experiments on Ethnic Discrimination in Rental Housing Mar-

kets," *Journal of Ethnic and Migration Studies* 45 (1) (2019): 95–114, https://doi.org/10.1080/1369183X.2018.1489223; and Lincoln Quillian, Devah Pager, Ole Hexel, and Arnfinn H. Midtbøen, "Meta-Analysis of Field Experiments Shows No Change in Racial Discrimination in Hiring over Time," *Proceedings of the National Academy of Sciences* 114 (41) (2017): 10870–10875, https://doi.org/10.1073/pnas.1706255114.

[8] Joe Feagin, *Systemic Racism: A Theory of Oppression* (Abingdon-on-Thames: Routledge, 2013); and Joe Feagin and Kimberley Ducey, *Racist America: Roots, Current Realities, and Future Reparations* (Abingdon-on-Thames: Taylor & Francis, 2000).

[9] Brian A. Nosek, Frederick L. Smyth, Jeffrey J. Hansen, et al., "Pervasiveness and Correlates of Implicit Attitudes and Stereotypes," *European Review of Social Psychology* 18 (1) (2007): 36–88, https://doi.org/10.1080/10463280701489053; and Kate A. Ratliff, Jennifer Howell, Colin Smith, et al., "Documenting Bias from 2007–2015," unpublished manuscript, last updated February 28, 2020, https://osf.io/rfzhu (accessed November 16, 2023).

[10] Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz, "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test," *Journal of Personality and Social Psychology* 74 (6) (1998): 1464–1480, https://doi.org/10.1037/0022-3514.74.6.1464.

[11] Richard E. Petty, Zakary L. Tormala, Pablo Brinol, and W. Blair G. Jarvis, "Implicit Ambivalence from Attitude Change: An Exploration of the PAST Model," *Journal of Personality and Social Psychology* 90 (1) (2006): 21–41, https://doi.org/10.1037/0022-3514.90.1.21; Laurie A. Rudman, "Sources of Implicit Attitudes," *Current Directions in Psychological Science* 13 (2) (2004): 79–82; and Timothy D. Wilson, Samuel Lindsey, and Tonya Y. Schooler, "A Model of Dual Attitudes," *Psychological Review* 107 (1) (2000): 101–126, https://doi.org/10.1037/0033-295X.107.1.101.

[12] Greenwald, McGhee, and Schwartz, "Measuring Individual Differences in Implicit Cognition"; and B. Keith Payne, "Conceptualizing Control in Social Cognition: How Executive Functioning Modulates the Expression of Automatic Stereotyping," *Journal of Personality and Social Psychology* 89 (4) (2005): 488–503, https://doi.org/10.1037/0022-3514.89.4.488.

[13] Erin Cooley and B. Keith Payne, "Using Groups to Measure Intergroup Prejudice," *Personality and Social Psychology Bulletin* 43 (1) (2017): 46–59, https://doi.org/10.1177/0146167216675331; Bertram Gawronski, Mike Morrison, Curtis E. Phills, and Silvia Galdi, "Temporal Stability of Implicit and Explicit Measures: A Longitudinal Analysis," *Personality and Social Psychology Bulletin* 43 (3) (2017): 300–312, https://doi.org/10.1177/0146167216684131; Wilhelm Hofmann, Jan De Houwer, Marco Perugini, Frank Baeyens, and Geert Crombez, "Evaluative Conditioning in Humans: A Meta-Analysis," *Psychological Bulletin* 136 (3) (2010): 390–421, https://pubmed.ncbi.nlm.nih.gov/20438144; and Keith Payne and Kristjen Lundberg, "The Affect Misattribution Procedure: Ten Years of Evidence on Reliability, Validity, and Mechanisms," *Social and Personality Psychology Compass* 8 (12) (2014): 672–686, https://doi.org/10.1111/spc3.12148.

[14] Andrew Scott Baron and Mahzarin R. Banaji, "The Development of Implicit Attitudes: Evidence of Race Evaluations from Ages 6 and 10 and Adulthood," *Psychological Science* 17 (1) (2006): 53–58, https://doi.org/10.1111/j.1467-9280.2005.01664.x; Yarrow Dunham, Eva E. Chen, and Mahzarin R. Banaji, "Two Signatures of Implicit Intergroup Attitudes: Developmental Invariance and Early Enculturation," *Psychological Science* 24 (6) (2013): 860–868, https://doi.org/10.1177/0956797612463081; and Brandon D. Stewart,

William Von Hippel, and Gabriel A. Radvansky, "Age, Race, and Implicit Prejudice: Using Process Dissociation to Separate the Underlying Components," *Psychological Science* 20 (2) (2009): 164–168, https://doi.org/10.1111/j.1467-9280.2009.02274.x.

[15] Anthony G. Greenwald, T. Andrew Poehlman, Eric Luis Uhlmann, and Mahzarin R. Banaji, "Understanding and Using the Implicit Association Test: III. Meta-Analysis of Predictive Validity," *Journal of Personality and Social Psychology* 97 (1) (2009): 17–41, https://doi.org/10.1037/a0015575; and Frederick L. Oswald, Gregory Mitchell, Hart Blanton, et al., "Predicting Ethnic and Racial Discrimination: A Meta-Analysis of IAT Criterion Studies," *Journal of Personality and Social Psychology* 105 (2) (2013): 171–192.

[16] Irene V. Blair and Elizabeth Brondolo, "Moving Beyond the Individual: Community-Level Prejudice and Health," *Social Science & Medicine* 183 (2017): 169–172, https://doi.org/10.1016/j.socscimed.2017.04.041; Eric Hehman, Jessica K. Flake, and Jimmy Calanchini, "Disproportionate Use of Lethal Force in Policing Is Associated With Regional Racial Biases of Residents," *Social Psychological and Personality Science* 9 (4) (2018): 393–401, https://doi.org/10.1177/1948550617711229; James R. Rae, Anna Kaisa Newheiser, and Kristina R. Olson, "Exposure to Racial Out-Groups and Implicit Race Bias in the United States," *Social Psychological and Personality Science* 6 (5) (2015): 535–543; and Brian A. Nosek, Frederick L. Smyth, Natarajan Sriram, et al., "National Differences in Gender-Science Stereotypes Predict National Sex Differences in Science and Math Achievement," *Proceedings of the National Academy of Sciences* 106 (26) (2009): 10593–10597, https://doi.org/10.1073/pnas.0809921106.

[17] B. Keith Payne, Heidi A. Vuletich, and Kristjen B. Lundberg, "The Bias of Crowds: How Implicit Bias Bridges Personal and Systemic Prejudice," *Psychological Inquiry* 28 (4) (2017): 233–248, https://doi.org/10.1080/1047840X.2017.1335568.

[18] Peter J. Rentfrow, Samuel D. Gosling, Markus Jokela, et al., "Divided We Stand: Three Psychological Regions of the United States and Their Political, Economic, Social, and Health Correlates," *Journal of Personality and Social Psychology* 105 (6) (2013): 996–1012, https://doi.org/10.1037/a0034434.

[19] Peter J. Rentfrow, Markus Jokela, and Michael E. Lamb, "Regional Personality Differences in Great Britain," *PLOS ONE* 10 (3) (2015), https://doi.org/10.1371/journal.pone.0122245.

[20] Thomas J. Allen, Jeffrey W. Sherman, and Karl Christoph Klauer, "Social Context and the Self-Regulation of Implicit Bias," *Group Processes and Intergroup Relations* 13 (2) (2010): 137–149, https://doi.org/10.1177/1368430209353635; Jamie Barden, William W. Maddux, Richard E. Petty, and Marilynn B. Brewer, "Contextual Moderation of Racial Bias: The Impact of Social Roles on Controlled and Automatically Activated Attitudes," *Journal of Personality and Social Psychology* 87 (1) (2004): 5–22, https://doi.org/10.1037/0022-3514.87.1.5; and Brian S. Lowery, Curtis D. Hardin, and Stacey Sinclair, "Social Influence Effects on Automatic Racial Prejudice," *Journal of Personality and Social Psychology* 81 (5) (2001): 842–855, https://doi.org/10.1037//0022-3514.81.5.842.

[21] Barden, Maddux, Petty, and Brewer, "Contextual Moderation of Racial Bias"; and Bernd Wittenbrink, Charles M. Judd, and Bernadette Park, "Spontaneous Prejudice in Context: Variability in Automatically Activated Attitudes," *Journal of Personality and Social Psychology* 81 (5) (2001): 815–827, http://doi.org/10.1037/0022-3514.81.5.815.

[22] Andrea R. French, Timothy M. Franz, Laura L. Phelan, and Bruce E. Blaine, "Reducing Muslim/Arab Stereotypes through Evaluative Conditioning," *The Journal of Social Psychology* 153 (1) (2013): 6–9, https://doi.org/10.1080/00224545.2012.706242.

[23] Hofmann, De Houwer, Perugini, et al., "Evaluative Conditioning in Humans."

[24] Wittenbrink, Judd, and Park, "Spontaneous Prejudice in Context"; and Nilanjana Dasgupta and Anthony G. Greenwald, "On the Malleability of Automatic Attitudes: Combating Automatic Prejudice with Images of Admired and Disliked Individuals," *Journal of Personality and Social Psychology* 81 (5) (2001): 800–814, https://doi.org/10.1037/0022-3514.81.5.800.

[25] Mary Mederios Kent, "Immigration and America's Black Population," *Population Bulletin* 62 (4) (2007), https://u.demog.berkeley.edu/~jrw/Biblio/Eprints/PRB/files/62.4immigration.pdf.

[26] Samantha Longman-Mills, Carole Mitchell, and Wendel Abel, "The Psychological Trauma of Slavery: The Jamaican Case Study," *Social and Economic Studies* 68 (3–4) (2019): 79–102.

[27] Avidit Acharya, Matthew Blackwell, and Maya Sen, "The Political Legacy of American Slavery," *The Journal of Politics* 78 (3) (2016): 621–641; and Nathan Nunn and Leonard Wantchekon, "The Slave Trade and the Origins of Mistrust in Africa," *American Economic Review* 101 (7) (2011): 3221–3252, https://doi.org/10.1257/aer.101.7.3221.

[28] Ellora Derenoncourt, Chi Hyun Kim, Moritz Kuhn, and Moritz Schularick, "Wealth of Two Nations: The U.S. Racial Wealth Gap, 1860–2020," NBER Working Paper 30101 (Cambridge, Mass.: National Bureau of Economic Research, 2022).

[29] Nancy Krieger, Jarvis T. Chen, Brent A. Coull, et al., "Jim Crow and Premature Mortality among the U.S. Black and White Population, 1960–2009: An Age-Period-Cohort Analysis," *Epidemiology* 25 (4) (2014): 494–504, https://doi.org/10.1097/EDE.0000000000000104.

[30] Jhacova Williams, "Historical Lynchings and the Contemporary Voting Behavior of Blacks," *American Economic Journal: Applied Economics* 14 (3) (2022): 224–253, https://doi.org/10.1257/app.20190549.

[31] Heather A. O'Connell, "Historical Shadows: The Links Between Sundown Towns and Contemporary Black–White Inequality," *Sociology of Race and Ethnicity* 5 (3) (2019): 311–325, https://doi.org/10.1177/2332649218761979.

[32] Sonali Bose, Jaime Madrigano, and Nadia N. Hansel, "When Health Disparities Hit Home: Redlining Practices, Air Pollution, and Asthma," *American Journal of Respiratory and Critical Care Medicine* 206 (7) (2022): 803–804, https://doi.org/10.1164/rccm.202206-1063ed; Clese E. Erikson, Randl B. Dent, Yoon Hong Park, and Qian Luo, "Historic Redlining and Contemporary Behavioral Health Workforce Disparities," *JAMA Network Open* 5 (4) (2022), https://doi.org/10.1001/jamanetworkopen.2022.9494; Nick Graetz and Michael Esposito, "Historical Redlining and Contemporary Racial Disparities in Neighborhood Life Expectancy," *Social Forces* 102 (1) (2023): 1–22, https://doi.org/10.1093/sf/soac114; Min Li and Faxi Yuan, "Historical Redlining and Resident Exposure to COVID-19: A Study of New York City," *Race and Social Problems* 14 (2) (2022): 85–100, https://doi.org/10.1007%2Fs12552-021-09338-z; Min Li and Faxi Yuan, "Historical Redlining and Food Environments: A Study of 102 Urban Areas in the United States," *Health & Place* 75 (2022): 102775, https://doi.org/10.1016/j.healthplace.2022.102775; Jeffrey Mitchell and Guilherme Kenji Chihaya, "Tract Level Associations Between Historical Residential Redlining and Contemporary Fatal Encounters with Police," *Social Science & Medicine* 302 (2022): 114989, https://doi.org/10.1016/j.socscimed.2022.114989; Issam Motairek, Zhuo Chen, Mohamed H. E. Makhlouf, et al., "Historical Neighbour-

hood Redlining and Contemporary Environmental Racism," *Local Environment* 28 (4) (2023): 518–528; Issam Motairek, Eun Kyung Lee, Scott Janus, et al., "Historical Neighborhood Redlining and Contemporary Cardiometabolic Risk," *Journal of the American College of Cardiology* 80 (2) (2022): 171–175, https://doi.org/10.1016/j.jacc.2022.05.010; Richard Casey Sadler, Thomas Walter Wojciechowski, Pamela Trangenstein, et al., "Linking Historical Discriminatory Housing Patterns to the Contemporary Alcohol Environment," *Applied Spatial Analysis and Policy* 16 (2) (2023): 561–581, https://doi.org/10.1007/s12061-022-09493-9; Leah H. Schinasi, Chahita Kanungo, Zachary Christman, et al., "Associations between Historical Redlining and Present-Day Heat Vulnerability Housing and Land Cover Characteristics in Philadelphia, PA," *Journal of Urban Health* 99 (1) (2022): 134–145, https://doi.org/10.1007/s11524-021-00602-6; Elli Schwartz, Nathaniel Onnen, Peter F. Craigmile, and Megan E. Roberts, "The Legacy of Redlining: Associations between Historical Neighborhood Mapping and Contemporary Tobacco Retailer Density in Ohio," *Health & Place* 68 (2021): 102529, https://doi.org/10.1016/j.healthplace.2021.102529; and Mudia Uzzi, Kyle T. Aune, Lea Marineau, et al., "An Intersectional Analysis of Historical and Contemporary Structural Racism on Non-Fatal Shootings in Baltimore, Maryland," *Injury Prevention* 29 (1) (2023): 85–90, http://doi.org/10.1136/ip-2022-044700.

[33] B. Keith Payne, Heidi A. Vuletich, and Jazmin L. Brown-Iannuzzi, "Historical Roots of Implicit Bias in Slavery," *Proceedings of the National Academy of Sciences* 116 (24) (2019): 11693–11698, https://doi.org/10.1073/pnas.1818816116.

[34] Jacob Orchard and Joseph Price, "County-Level Racial Prejudice and the Black-White Gap in Infant Health Outcomes," *Social Science & Medicine* 100 (181) (2017): 191–198, https://doi.org/10.1016/j.socscimed.2017.03.036.

[35] Marilyn D. Thomas, Eli K. Michaels, Sean Darling-Hammond, et al., "Whites' County-Level Racial Bias, COVID-19 Rates, and Racial Inequities in the United States," *International Journal of Environmental Research and Public Health* 17 (22) (2020): E8695, https://doi.org/10.3390/ijerph17228695.

[36] Eli K. Michaels, Christine Board, Mahasin S. Mujahid, et al., "Area-Level Racial Prejudice and Health: A Systematic Review," *Health Psychology* 41 (3) (2022): 211–224, https://doi.org/10.1037%2Fhea0001141; and Tessa E. S. Charlesworth and Mahzarin R. Banaji, "Evidence of Covariation Between Regional Implicit Bias and Socially Significant Outcomes in Healthcare, Education, and Law Enforcement," in *Handbook on Economics of Discrimination and Affirmative Action*, ed. Ashwini Deshpande (Berlin: Springer Nature, 2023), 593–613.

[37] Mark J. Chin, David M. Quinn, Tasminda K. Dhaliwal, and Virginia S. Lovison, "Bias in the Air: A Nationwide Exploration of Teachers' Implicit Racial Attitudes, Aggregate Bias, and Student Outcomes," *Educational Researcher* 49 (8) (2020): 566–578, https://doi.org/10.3102/0013189X20937240.

[38] Francis A. Pearman and Ebony O. McGee, "Anti-Blackness and Racial Disproportionality in Gifted Education," *Exceptional Children* 88 (4) (2022): 359–380.

[39] Sarah Beth Bell, Rachel Farr, Eugene Ofosu, Eric Hehman, and C. Nathan DeWall, "Implicit Bias Predicts Less Willingness and Less Frequent Adoption of Black Children More Than Explicit Bias," *The Journal of Social Psychology* 163 (4) (2023): 554–565.

[40] Pierce D. Ekstrom, Joel M. Le Forestier, and Calvin K. Lai, "Racial Demographics Explain the Link between Racial Disparities in Traffic Stops and County-Level Racial Attitudes," *Psychological Science* 33 (4) (2022): 497–509, https://doi.org/10.1177/09567976211053573;

and Marleen Stelter, Iniobong Essien, Carsten Sander, and Juliane Degner, "Racial Bias in Police Traffic Stops: White Residents' County-Level Prejudice and Stereotypes Are Related to Disproportionate Stopping of Black Drivers," *Psychological Science* 33 (4) (2022): 483–496, https://doi.org/10.1177/09567976211051272.

[41] Hehman, Flake, and Calanchini, "Disproportionate Use of Lethal Force in Policing Is Associated With Regional Racial Biases of Residents."

[42] Tyler Jimenez, Peter J. Helm, and Jamie Arndt, "Racial Prejudice Predicts Police Militarization," *Psychological Science* 33 (12) (2022): 2009–2026, https://doi.org/10.1177/09567976221112936.

[43] B. Keith Payne and Julian M. Rucker, "Explaining the Spatial Patterning of Racial Disparities in Traffic Stops Requires a Structural Perspective: Further Reflections on Stelter et al. (2022) and Ekstrom et al. (2022)," *Psychological Science* 33 (4) (2022): 666–668, https://doi.org/10.1177/09567976211056641.

[44] Mimi V. Chapman, William J. Hall, Kent Lee, et al., "Making a Difference in Medical Trainees' Attitudes toward Latino Patients: A Pilot Study of an Intervention to Modify Implicit and Explicit Attitudes," *Social Science & Medicine* 199 (2018): 202–208, https://doi.org/10.1016%2Fj.socscimed.2017.05.013; Dasgupta and Greenwald, "On the Malleability of Automatic Attitudes"; Alexander J. Stell and Tom Farsides, "Brief Loving-Kindness Meditation Reduces Racial Bias, Mediated by Positive Other-Regarding Emotions," *Motivation and Emotion* 40 (1) (2016): 140–147, https://doi.org/10.1007/s11031-015-9514-x; and Andrew R. Todd, Galen V. Bodenhausen, Jennifer A. Richeson, and Adam D. Galinsky, "Perspective Taking Combats Automatic Expressions of Racial Bias," *Journal of Personality and Social Psychology* 100 (6) (2011): 1027–1042.

[45] Patrick S. Forscher, Calvin K. Lai, Jordan R. Axt, et al., "A Meta-Analysis of Procedures to Change Implicit Measures," *Journal of Personality and Social Psychology* 117 (3) (2019): 522–559, https://doi.org/10.1037/pspa0000160; and Calvin K. Lai, Allison L. Skinner, Erin Cooley, et al., "Reducing Implicit Racial Preferences: II. Intervention Effectiveness Across Time," *Journal of Experimental Psychology: General* 145 (8) (2016): 1001–1016, https://doi.org/10.1037/xge0000179.

[46] B. Keith Payne and Heidi A. Vuletich, "Policy Insights from Advances in Implicit Bias Research," *Policy Insights from the Behavioral and Brain Sciences* 5 (1) (2018): 49–56, https://doi.org/10.1177/2372732217746190.

[47] Derenoncourt, Kim, Kuhn, and Schularick, *Wealth of Two Nations*.

[48] George B. Cunningham, "The Influence of Group Diversity on Intergroup Bias following Recategorization," *The Journal of Social Psychology* 146 (5) (2006): 533–547.

[49] Eric Luis Uhlmann and Geoffrey L. Cohen, "Constructed Criteria: Redefining Merit to Justify Discrimination," *Psychological Science* 16 (6) (2005): 474–480.

[50] Dawn Culpepper, Damani White-Lewis, KerryAnn O'Meara, Lindsey Templeton, and Julia Anderson, "Do Rubrics Live Up to Their Promise? Examining How Rubrics Mitigate Bias in Faculty Hiring," *The Journal of Higher Education* 94 (7) (2023): 823–850, https://doi.org/10.1080/00221546.2023.2168411.

[51] Thomas E. Ford, Frank Gambino, Hanjoon Lee, et al., "The Role of Accountability in Suppressing Managers' Preinterview Bias against African-American Sales Job Applicants," *Journal of Personal Selling & Sales Management* 24 (2) (2004): 113–124.

[52] Payne, "Conceptualizing Control in Social Cognition"; B. Keith Payne, "Weapon Bias: Split-Second Decisions and Unintended Stereotyping," *Current Directions in Psychological Science* 15 (6) (2006): 287–291, https://doi.org/10.1111/j.1467-8721.2006.00454.x; and Jeffrey W. Sherman, Bertram Gawronski, Karen Gonsalkorale, et al., "The Self-Regulation of Automatic Associations and Behavioral Impulses," *Psychological Review* 115 (2) (2008): 314–335, https://doi.org/10.1037/0033-295x.115.2.314.

[53] Simon Calmar Andersen and Thorbjørn Sejr Guul, "Reducing Minority Discrimination at the Front Line–Combined Survey and Field Experimental Evidence," *Journal of Public Administration Research and Theory* 29 (3) (2019): 429–444, https://doi.org/10.1093/jopart/muy083.

[54] Sophie Trawalter, D-J Bart-Plange, and Kelly M. Hoffman, "A Socioecological Psychology of Racism: Making Structures and History More Visible," *Current Opinion in Psychology* 32 (2020): 47–51, https://doi.org/10.1016/j.copsyc.2019.06.029.

"When the Cruiser Lights Come On": Using the Science of Bias & Culture to Combat Racial Disparities in Policing

Rebecca C. Hetey, MarYam G. Hamedani, Hazel Rose Markus & Jennifer L. Eberhardt

In this essay, we highlight the interplay between individuals' psychological processes and sociocultural systems in producing and maintaining racial bias. We use a conceptual tool we call the culture cycle to map these dynamics, and illustrate them with research and in-depth examples from our work reducing racial disparities in routine policing in Oakland, California. We feature the most common police encounter – the vehicle stop – and highlight evidence-based interventions we developed both to reduce the frequency of vehicle stops and mitigate racial disparities in stops. Throughout, we draw on our expertise in the social psychology of bias, culture, and inequality, as well as our experiences building research-driven partnerships with public- and private-sector leaders, to inform organizational and societal change.

It was 1999. By almost anyone's account, crime was out of control in Oakland, California. And it seemed as though Oakland police officers would stop at nothing to curb it. Innocent residents described cops slipping drugs into their pockets or purses. A woman was forced to strip naked in the street, as one officer searched her and another planted drugs in the trunk of her car. A father, who was taking his son to his first visit to a barber shop, had his nose broken and teeth knocked loose by a cop. A man who made the mistake of double parking his car was beaten. The list went on and on.[1] Officers were on a mission to "handle" anyone who looked like they were prone to make trouble. It was the beginning of a pattern that would last for years.

Civil rights attorneys John Burris and Jim Chanin were used to Oakland residents coming to their offices with stories about police misconduct and brutality. By the summer of 2000, these stories increasingly featured the same group of Oakland officers working in the same part of the city and using the same violent and illegal tactics.[2] Francisco "Choker" Vazquez was the ringleader of this group of vigilante cops who called themselves "the Riders." They worked the night shift

https://doi.org/10.1162/daed_a_02052

"ruling their beat with an iron fist."[3] They allegedly assaulted, kidnapped, planted evidence on, and filed false police reports against their victims, almost all of whom were Black.

Burris and Chanin had spent their careers fighting the mistreatment of Black people at the hands of police. The attorneys dismissed the most immediately obvious explanation that the Riders were a "few bad apples" – four racist, violent officers out of a force of more than seven hundred – who had gone rogue. They considered the Riders' misconduct to be symptomatic of a larger, systemic problem. And so, instead of suing the four officers, Burris and Chanin filed a claim against the City of Oakland that named dozens of members of the Oakland Police Department (OPD), including the chief of police, and alleged that cover-ups and poor supervision allowed such egregious misconduct to happen with impunity.[4]

In the words of former Oakland police captain Ronald Davis, the emergence of the Riders in the late 1990s was "completely predictable." Davis, who spent twenty years with the OPD and would later go on to lead Barack Obama's President's Task Force on 21st Century Policing, described the Riders as "part of this culture [at the OPD]. They didn't come out of nowhere. . . . Everything was tied to it from the leadership, to the messaging, to the strategies, to the tactics, to the lack of accountability."[5] Burris and Chanin's claim was their first step toward trying to bring about meaningful change to the department as a whole.

Within the OPD, a number of organizational levers could have been pulled to learn about rogue cops and improve the nature of police-community interactions. Stakeholders in Oakland began to pull some of those levers for change over the next two decades. The catalyst was Burris and Chanin filing a civil rights lawsuit against the City of Oakland, arguing that the OPD had engaged in a sustained pattern and practice of denying Black Oaklanders' civil rights and would need to eliminate the toxic aspects of its police culture.[6] Some of these cultural features were particular to the OPD, but others reflected more widespread issues in the profession and its troubled history with Black communities.[7] Armed with the science of racial bias and culture, we later joined their quest – not by marshaling the justice system, but by using data to spur change.

Racial disparities and bias are not static properties of institutions and organizations that can be found and extracted. Rather disparities can lead to bias, and bias can lead to disparities in a mutually reinforcing process. Racial bias is deeply ingrained in the architecture of our minds and woven throughout all facets of our society: our history and narratives; our institutions, laws, and policies; our norms and practices; our interpersonal interactions; and our psychology and actions.[8] In other words, our culture.

Culture can be broadly defined as a socially meaningful system of shared ideas, histories, policies, practices, norms, and products that structure and organize behavior.[9] Conceptualizing implicit racial bias as merely a byproduct of human

cognition overlooks the critical scientific insight that racial bias exists not only in the head, but also in the world. Implicit bias is the residue that an unequal world leaves on an individual's mind and brain, residue that has been created and built into institutional policies and practices and socialized into patterns of behavior over hundreds of years through the workings of culture.[10] After decades of a narrower cognitive approach to bias, a broader, more systemic, multilevel perspective is having a rebirth in social psychology: what we call a *sociocultural approach* to racial bias.

To help make that sociocultural approach more concrete, we developed and have long used a conceptual tool called the *culture cycle*,[11] which can map the complex, dynamic interplay between racial bias as an individual-level phenomenon and the systemic ways in which bias might operate at other levels of culture. More specifically, the culture cycle contains four levels of culture – ideas, institutions, interactions, and individuals – and each level dynamically influences and interacts with the other levels. As such, the culture cycle can help researchers and practitioners alike identify where bias lurks and how it manifests in different settings. It can also be used to diagnose which features of an institution or organization produce and maintain bias, and prescribe how those features can be altered to mitigate bias and to reduce racial disparities.[12]

For more than a decade, we have been using this tool and approach in the field.[13] We and our collaborators have worked with a variety of stakeholders in the criminal justice system to identify, unpack, and address racial disparities and the bias that can spring from them.[14] Through in-depth analyses of law enforcement policies and procedures, as well as actual police-community encounters, we have used this approach to reduce racial disparities in policing in particular. We start by focusing on the most familiar police encounter: the vehicle stop. Though the public's attention on matters of racial justice and policing often centers on the fatal use of excessive force, everyday police stops are in fact the most common point of contact by which members of the public meet – and in some ways collide with – the institution of policing.[15] The police stop nearly 18.7 million drivers each year in the United States, yet not all racial groups have the same experience during these stops.[16] Black drivers are not only more likely to be stopped than any other racial demographic, they are also more likely to be searched, handcuffed, and arrested.[17] And they experience such outcomes at an elevated rate, despite the fact that they are significantly more likely than white drivers to be stopped for discretionary reasons that have little to do with public safety (such as having incorrectly displayed license plates).[18]

Our work excavating the culture of policing began in Oakland. The OPD had been plagued by scandals for decades.[19] In fact, the four officers engaged in the Riders' scandal were highly respected in the department,

shaping the very idea of what it meant to be an effective cop in the city and contributing to an aggressive culture of policing.[20] It was a rookie officer with just nine shifts on the job who became the internal whistleblower, setting off an investigation and ultimately shutting down the operation.[21] As brand new to the OPD, he could recognize the Riders' actions as violating everything he had just learned in the police academy. The four officers were fired and charged with forty-eight felonies, but not one of them was ever convicted.

The whistleblower's claims going public and the "high-visibility arrest of the Riders was really the impetus for clueing us in on this," Burris recalled.[22] Burris and Chanin began tracking down the Riders' victims and building a case that highlighted the systemic issues at the OPD that were in need of change. In December of 2000, Burris and Chanin filed a class-action lawsuit against the City of Oakland on behalf of 119 plaintiffs, 118 of whom were Black.[23] Collectively, they spent over forty years (14,665 days) in prison for crimes they did not commit.[24] The lawsuit eventually led to a $10.9 million settlement for the plaintiffs, mandatory federal oversight of the OPD, and a series of more than four dozen reforms required of the agency.

The oversight agreement required the OPD to collect data on its routine police stops by race. Yet it took nearly ten years for the department to collect reliable data. In the spring of 2014, the plaintiffs' attorneys and the federal monitor asked Jennifer Eberhardt to serve as a subject matter expert. She was brought in to analyze the department's stop data, determine whether there were significant racial disparities, and suggest ways to improve police-community interactions. Burris and Chanin made it clear that what they really wanted to know was "What happens when the cruiser lights come on?" That is, why are so many Black people stopped by police in Oakland? How do officers approach them? And how do those interactions unfold?

After assembling an initial team, including Rebecca Hetey and Benoît Monin, a fellow social psychologist and colleague at Stanford, our first step was to learn how to navigate the broader context and to learn the roles of the people within it.

While a vehicle stop may at first appear to be an interaction between two individuals, on-duty law enforcement officers are in fact acting in their capacity as representatives of a powerful institution and the government itself.[25] As institutional actors, individual police officers are embedded in complex power dynamics and are bound up by systems and subject to policies, practices, and laws that could put them in a position to produce and reproduce racial disparities and systemic inequity, all without their needing to personally endorse racial stereotypes or inequality or even be aware of the broader impact their actions could potentially have.[26] Police officers are part of a hierarchical, highly interdependent, paramilitary organization with strong social norms and rigid expectations, if not explicit policies, that dictate nearly all aspects of their behavior.[27] Status is accorded based on years of experience and rank. Deference to those higher up the chain of command is re-

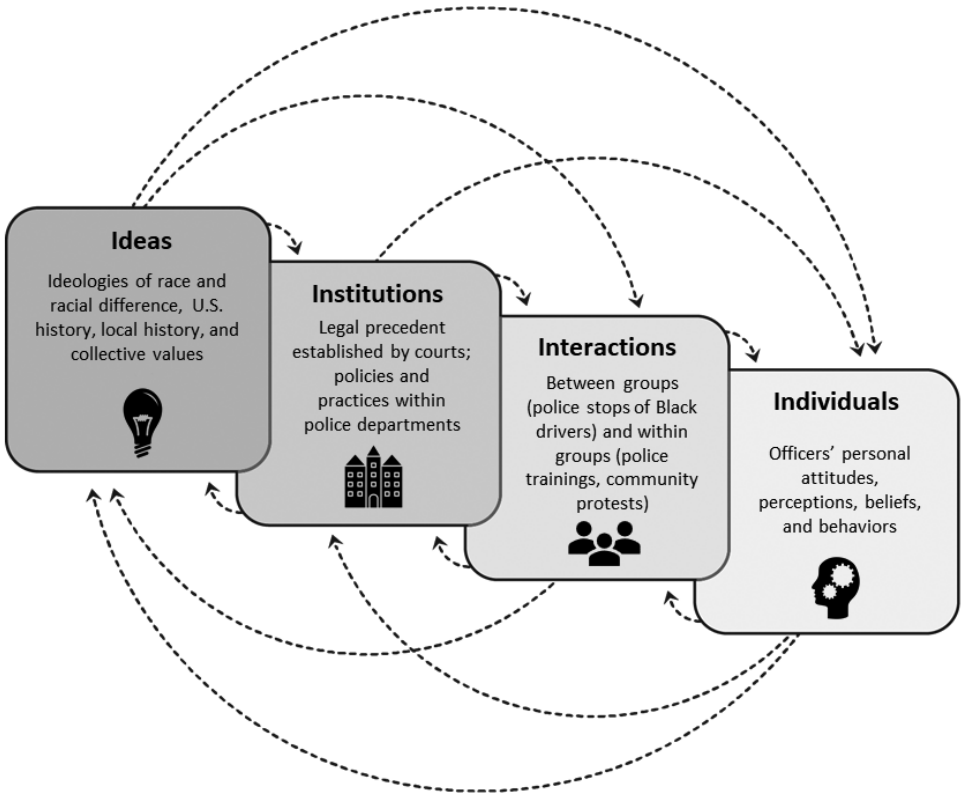
quired, and loyalty to peers within the same rank is of supreme importance.[28] The Riders' illegal tactics, for instance, were open secrets at the OPD.

Aided by the culture cycle (Figure 1),[29] our first step was to map the multilevel dynamics at play. We interviewed and spoke with a diverse set of individuals to take the pulse of police-community relations in Oakland, to be better positioned to diagnose the problems we learned about, and to see how they manifested in the local context. From the police chief to the rank and file, from the mayor to the city council, from the federal judge to the federal monitor appointed to provide oversight, from the plaintiffs' attorneys to the residents of Oakland, many different perspectives are activated when the cruiser lights come on. Interactions between police officers and community members that take place during the vehicle stop are guided, in part, by relevant laws, policies, and practices, and can lay the foundation for an entire community's relationship with police. Each encounter holds the potential to significantly increase or decrease public trust, to become the site of police violence that can set off racial unrest, or everything in between. Our focus on the vehicle stop was sadly prescient: George Floyd was killed after a Minneapolis police officer forcibly removed him from his car, igniting one of the largest public mobilizations in U.S. history as people took to the streets to protest racial injustice.[30]

Initially, many people at the OPD expressed skepticism about our presence there; some displayed outright hostility. We were outsiders – outsiders from the reputedly liberal, elitist world of academia. They were convinced that we had set out only to uncover evidence that would prove our preexisting conclusion that they were all racists who engaged in deliberate racial profiling. Instead, the officers with whom we collaborated found that we were invested in this work for the long haul. We were driven to find strategies for improving policing and police-community relations. The rank and file discovered that we were genuinely interested in how they made sense of their jobs and how they saw themselves vis-à-vis the community members they encountered daily. We sought out opportunities to learn about their lived experiences, including going on ride-alongs (riding in the passenger seat of police vehicles to observe officers on patrol during their shift) at all hours of day and night, sitting in on trainings, delivering trainings, and witnessing how officers interact with one another. As we heard repeatedly and saw firsthand, there is an intense loyalty and interdependence among officers, forged in the knowledge that whether one would live to go home at the end of their shift could depend on a fellow officer's actions.[31] We had conversations with members of the OPD about why they had chosen to become police officers, what the worst elements of the job were, and what they wished the public knew.

At the same time, we engaged the community. We held meetings and hosted focus groups to learn about the OPD's enforcement practices directly from those who were impacted. Centering the voices of Oakland community members in our work shed light on sources of tension between the police and the community. It

Figure 1
Applying the Culture Cycle to Map Racial Disparities and Bias in Law Enforcement



The culture cycle is a conceptual tool representing a sociocultural approach that can also be used to guide culture change. Here, we apply it to locate and address racial disparities and bias in law enforcement. All four levels of culture–ideas, institutions, interactions, and individuals–dynamically interact and influence one another (as indicated by the cycling arrows) and are equally important (all four boxes are the same size). Source: Image adapted from Alan P. Fiske, Shinobu Kitayama, Hazel Rose Markus, and Richard E. Nisbett, "The Cultural Matrix of Social Psychology," in *The Handbook of Social Psychology*, ed. Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey (New York: McGraw Hill, 1998), 915–981; MarYam G. Hamedani, Hazel Rose Markus, Rebecca C. Hetey, and Jennifer L. Eberhardt, "We Built This Culture (So We Can Change It): Seven Principles for Intentional Culture Change," *American Psychologist* (advance online publication, 2023), https://doi.org/10.1037/amp0001209; Hazel Rose Markus and Alana Conner, *Clash!: How to Thrive in a Multicultural World* (New York: Plume, 2014); and Hazel Rose Markus and Shinobu Kitayama, "Cultures and Selves: A Cycle of Mutual Constitution," *Perspectives on Psychological Science* 5 (4) (2010): 420–430.

was during these meetings that we heard about the central role that respect plays in police stops.[32] And as a result, we focused on the issue of respect in our own research. Black men, in particular, described their concerns about being stopped and assumed to be criminal. They spoke not only about the lack of respect they received during those stops, but also described being handcuffed for minor infractions like driving with expired registration tags – as officers fished for a reason to arrest them.

Building relationships with community members also came with challenges. Some people felt the problems were intractable and the department too broken to be salvaged. We heard stories about the physical and emotional scars left by the OPD's often brutal history. We read about the political pressure in Oakland in the late 1990s to hire more police and adopt aggressive crime fighting tactics as part of the mayor's "campaign against crime and grime," and how this climate helped give rise to the Riders.[33]

Appreciating the role of history is integral to this work. We observed current officers watch a documentary about the OPD from 1974 that aired on a local Bay Area television station at the time.[34] The narrator explained: "During the fifties, Oakland became a stagnant, seething ghetto of impoverished Blacks and Chicanos surrounded by the white affluence of the Oakland Hills.... To contain the misery and violence of the ghetto, Oakland's all-white police department earned a reputation for head-knocking brutality that has left a well-remembered legacy of bitterness in the minds and hearts of many who lived in that time and place." We saw the expressions on the current officers' faces, as they wrestled with the evidence right before their eyes that history can keep repeating itself, and asked what that meant for them and how it affected the ways residents still view them. As many have noted, the very origins of policing as an institution in the United States can be traced back to slave patrols – and that history has been woven into the fabric of policing, whether the officers who wear the uniform today realize it or not.[35]

In the next phase of our work, we conducted a statistical analysis of twenty-eight thousand pedestrian and vehicle stops made by the OPD between 2013 and 2014 and found a consistent pattern of racial disparities across the entire course of the stop, from an officer's initial decision to stop a person to subsequent decisions to search, handcuff, and arrest that person.[36] The raw disparities were striking: roughly 60 percent of the stops officers made in Oakland were of Black people, although Black residents made up only 28 percent of Oakland's population at that time. Black people were disproportionately stopped even when we statistically accounted for two dozen factors, including the crime rate and the racial demographics of residents in the areas where the stops occurred. Further, once stopped, Black people were significantly more likely to be handcuffed, searched, and arrested – echoing a pattern of results found in cities across the country.[37] In fact, in Oakland we found that one in four Black men were handcuffed during the course of routine

stops (compared to only one in fifteen white men) – a statistic in complete alignment with what we heard directly from Black men themselves.

But simply uncovering the existence of racial disparities is not the same thing as understanding *where*, *why*, and *how* those disparities originated. What else is needed? Our approach as social psychologists was to rely on core principles that we know can disrupt and mitigate bias, once we had located both its situational triggers and the features of the sociocultural environment that perpetuate and sustain it.[38] We understood that change at the OPD, and for the policing industry more broadly, would have to be brought about purposefully and deliberately and it would require a supportive climate. We had to build relationships with those in power within the organization to put interventions in place that could alter the features of the OPD's culture that contribute to bias. Aided by the culture cycle, we probed for the sources of racial disparities across all levels of culture.

Although we found evidence for significant racial disparities at every point in the course of stops, many OPD officers pushed back. They believed our stop-data analysis to be incomplete, that "numbers can't tell the full story." They offered two common refrains: 1) "We don't racially profile, we *criminally* profile"; and 2) "If we don't make as many stops as we do now, crime will go up." We listened as one officer after another described how they often could not even see the race of the driver through the car window, making it virtually "impossible" to stop someone simply because they are Black. Yet we knew that the stop-data form they completed during stops included the question, "Could you determine the race/ethnicity of the individual(s) prior to the stop?" Among stops made in which the officer had reported not knowing the race of the driver prior to the stop, we found 48 percent of those stopped were Black. In contrast, among stops made in which the officer reported knowing the race of the driver, 62 percent of those stopped were Black.[39] In other cities, researchers have found that Black drivers are less likely to be stopped after sunset, when presumably officers are less able to see the race of the driver.[40] Nevertheless, officers insisted that the "vast majority" of the stops that the OPD made were, in fact, based on previous intelligence. Here, intelligence refers to information, such as suspect descriptions provided by crime victims or specific patterns of gang activity or illegal drug dealing, as opposed to relying on intuition. In other words: could the officer tie that particular person to a specific crime prior to the stop? We were told that since many of the drivers the OPD stopped were already on their radar for some reason, the stops served as a deterrent. And, they said, if the OPD did not make as many of these stops, crime would increase, and the community would not be happy.

Were officers simply using a straightforward crime reduction strategy? Or was their approach potentially laden with bias? When officers on patrol ask themselves, "Who should I pull over?" might cultural as-

sociations between Blackness and crime supply an answer? Bias can be embedded in our ideologies and our history, which in turn shape institutional and organizational policies and practices, which then influence interpersonal interactions, as well as individuals' attitudes, perceptions, beliefs, and behaviors.[41] An individual's behaviors can then either reinforce or disrupt those biases. Bias can also be situationally triggered.[42] In the realm of policing, we can see these larger cultural forces at play, specifically ideas and associations that link race and crime in the mind, institutional practices that prompt officers to rely on implicit stereotypes when judging suspicion and potential criminality, and interactions police have with each other and with the community that reify racial stereotypes.[43] Going deeper into the mechanics of the vehicle stop can help illustrate the situational nature of bias, as well as reveal ways to alter the situation to curb bias.

The Black-crime association can pervade all levels of law enforcement culture, just as it can pervade mainstream U.S. culture more generally.[44] This association has been passed down through history and persists today, promoted, for example, through social media posts by police departments that overrepresent Black suspects relative to local arrest rates.[45] Though not necessarily intentional, these posts nonetheless can amount to the government reinforcing racial associations between Blackness and crime. Implicit associations linking Blackness with crime, violence, and animals have been shown to be strong enough to alter individuals' basic perceptions in ways that further reinforce these stereotypes and serve to rationalize the harsh treatment of Black people.[46]

Experiments, for instance, have demonstrated that, in a mutually reinforcing way, exposure to Black people prompts thoughts of crime, and the concept of crime draws attention to Black people.[47] In one experiment, Eberhardt and colleagues subliminally primed police officers with the concept of crime by exposing them to words like "arrest" and "shoot" on a computer screen for mere fractions of a second, so fast that they could not say what they had seen. After being primed, officers became significantly more likely to attend to pictures of Black faces. When officers were asked directly, "Who looks criminal?" they chose more Black faces than white faces, and the more stereotypically Black the face, the more likely officers were to report that the face looked criminal. If the mere concept of crime causes police officers to be more vigilant to Black faces, what are the ramifications of being in a patrol car for ten- to twelve-hour shifts looking for criminals while hearing "male Black" continuously broadcast on the radio as fellow officers describe who has drawn their suspicion?[48]

These studies provide empirical evidence that pervasive stereotypes and associations linking race and crime at the ideas level of culture affect people's perceptions and actions at the individuals level; in policing, these biases become formalized and can have life and death consequences at the levels of institutions and interactions. Most apparent is the way stereotypes are acted out through the inter-

actions a police officer has with community members and the ways in which the officer may – intentionally or not, subtly or not – treat the Black and white civilians they encounter differently.

Officers also interact with each other. They watch their fellow officers disproportionately pull over Black people who are not engaged in any serious crime. They watch supervisors fail to address this behavior as problematic, or even praise officers who engage in it as "productive" and "proactive." Norms around what "we" (for example, police officers in this department) do and are supposed to do in a given situation powerfully shape behavior.[49] In the world of policing, norms influence routine enforcement practices by signaling to institutional actors what their own behavior could and should look like, separate from particular officers' individual motives and intentions. Moreover, research has shown that people go beyond conforming to norms to also making prescriptive leaps, justifying what we do as what we *should* do.[50] The result is that the underlying Black-crime association is reinforced.

Official policy, practice, and trainings encourage individual police officers to make stops simply to check people out and see who might be "up to no good," which also implicitly prompts officers to rely on stereotypes about race and crime.[51] In *Whren v. United States* (1996), the U.S. Supreme Court held that police officers can use traffic violations as a pretext to make a stop and investigate an unrelated crime for which they have little or no evidence.[52] The court ruled that, under the Fourth Amendment, once a police officer has probable cause or reasonable suspicion that a driver has committed a traffic violation, they can legally stop and detain the driver, regardless of what the officer's actual motive or reason for the stop may be. As such, the *Whren* decision effectively enshrined law enforcement's ability to act on their hunches about who is "up to no good."[53] Because violations can be minor and the underlying traffic law not normally enforced (for instance, driving less than five miles an hour over the speed limit), the majority of drivers likely commit one or more traffic infractions on a given day, meaning that law enforcement has broad discretion to detain people.[54] As such, so-called pretextual stops have been called "America's most egregious police practice" and some jurisdictions, including the State of Virginia in 2020, have moved to ban or limit their use.[55]

Race can shape the practice of routine police stops and routine police stops can shape ideas about race.[56] Baked into law enforcement exerting a larger footprint with Black communities, while engaging in underenforcement with white communities, are assumptions about who is more likely to have committed some offense and therefore "deserves" to be treated punitively. At play, too, are assumptions about who is "worthy" of being treated with compassion and given the benefit of the doubt. Americans have been shown to associate the concepts of payback and retribution with Black people, and the concepts of mercy and leniency with

white people.[57] More generally, people are willing to go the extra mile to help members of their in-groups, while simultaneously being more likely to harm members of out-groups.[58]

Community members who witness law enforcement's disparate interactions with the public likely leave with very different lessons depending on their own race and prior experiences with police. Every time a passerby observes a Black man being stopped by police and handcuffed on the side of the road, even when he has not committed a serious offense, another opportunity is presented for the Black-crime association to be strengthened. Indeed, exposure to such stark racial disparities in the criminal justice system has been shown to cause white people to become more supportive of harsh and punitive criminal justice policies that contribute to those disparities in the first place, further fueling the vicious cycle.[59] White community members might similarly become more likely to call the police on Black people for noncriminal activities (such as walking while Black), bringing police and members of marginalized communities into contact, and potentially putting them on a collision course. This occurs without any institutional actor from the justice system needing to initiate the encounter. Calling the police on Black people for mundane, noncriminal activities and/or making ill-informed claims of misconduct based on one's own biases has been termed "bias by proxy."[60] When police act on such calls, they risk being co-opted as an apparatus of other people's racial bias and being cast in the role of perpetuating the fear of Black people.

Black people not only report being disproportionately stopped by police, they witness each other having these interactions.[61] Racial disparities in such stops create additional interactions with a government institution that they feel regards Black people with suspicion and calls into question their status as equal citizens, free to move about without government intrusion and surveillance.[62] The vehicle stop, therefore, can strengthen the countervailing association that police are unfair and racist, which further affects how Black people will feel and interpret the actions of police the next time the cruiser lights come on.

Because of the ramifications of routine enforcement, we first worked to identify evidence-based strategies to, as our colleague Benoît Monin would say, "reduce the footprint of policing" in Oakland. With Monin playing a leading role, our research team worked with a task force of changemakers assembled by a deputy chief to reduce the number of stops of Black drivers who were not committing any serious crimes. We were prepared for officers to push back, to tell us that the "vast majority" of their stops were based on prior knowledge, or intelligence. For the most part, they would say, they stop Black drivers they can link to criminal activity. And because Black people are disproportionately linked to criminal activity, they are disproportionately stopped. When we asked what percent-

age constituted a "vast majority" of stops based on prior intelligence, we heard responses like 85 percent, 90 percent, even 99 percent. Our next question was what evidence supported their claim. The answer: None. "We don't track which specific stops are intelligence-led."

As an intervention strategy, we decided to leverage the decision-making process and officers' own understandings of *why* they were making these stops. The method we co-constructed with the OPD was simple on its face. Yet, while practical and easy to implement, it was grounded in the principles of social psychology. To decrease the likelihood that an officer's decision-making was being driven by their association between Blackness and crime, we would require all officers to ask themselves a specific question before each and every stop they considered making: *Is this stop intelligence-led?* We added this question to the stop-data form officers are required to fill out whenever they make a stop. If they indicated that the stop was intelligence-led, they had to list the specific source of that prior knowledge. This intervention would allow us to collect data on how often OPD officers were making stops based on prior intelligence, and empirically test whether objective reality matched officers' subjective reality.

Requiring officers to indicate whether each stop was intelligence-led is an intervention designed to mitigate specific situational triggers of bias and, in the process, alter the way officers make the decision to pull someone over. Responding quickly, relying on subjective standards, following cultural norms that do not challenge bias, a lack of accountability, and a lack of training have all been shown to exacerbate bias.[63] First, the intelligence-led question forced individual institutional actors to slow down in the moment and consider their reasons for making the contact. Second, directing their attention to previously collected intelligence encouraged officers to use more objective information rather than relying on their hunches and intuition about which drivers might be "up to no good," which we know can be tainted by racial bias. Third, by collecting and tracking the data and asking officers to document the specific source of intelligence, the question signaled that they were being monitored and could be held accountable for the nature of their stops. Fourth, OPD officers were trained on how to complete the form and what it meant for a stop to indeed be intelligence-led. What at first felt obvious to officers in fact necessitated discussion and guidelines to arrive at a consensus definition, which provided more clarity and explicit direction from supervisors. Finally, OPD leadership began to prioritize intelligence-led stops. The question on the form served as a salient, constant reminder. A seemingly small change helped shift broader norms in the department for what good policing looks like.

In the year before adding the intelligence-led question to the form, roughly thirty-two thousand people were stopped across the city; in the year after adding this question, the number dropped to less than twenty thousand. Stops of Black drivers in particular fell by 43 percent.[64] Again, a common refrain within

this context (that can block change) is that if police are not proactive and do not make as many stops, crime will increase. Yet as OPD officers made fewer stops, the crime rate in Oakland continued to decline. Reducing the number of stops, then, can both lessen the negative impact on people's lives and maintain public safety. These results pushed officers to rethink what they had always taken to be true and to make a change that was ultimately in their own best interests and also helped the agency prioritize its limited resources.[65]

What of officers' claim that the "vast majority" of stops were based on intelligence? Deciding on a standard definition of an intelligence-led stop and tracking the data showed that the percentage was not somewhere between 85 percent and 99 percent, as officers had maintained, but closer to 20 percent. This gap between what officers believed was happening and what was actually happening sparked another change in the agency. Officers experienced (perhaps for the first time) the benefit of collecting data. This one question – is this stop intelligence-led? – and the ensuing data collection provoked conversations. It pushed both police executives and the rank and file to be more reflective. It led them to a deeper understanding of issues of race and policing. As another sign of broader culture change at the agency, the OPD now routinely questions how race could influence their decision-making and seeks out data to inform the development of policy and practice, one of fifty recommendations we made to the OPD and other agencies to mitigate racial disparities and improve police-community relations.[66]

In addition to working to reduce the likelihood that Oaklanders would be stopped, we also intervened so stops that did occur would proceed more respectfully. We leveraged officers' body-worn camera footage to better understand the nature of police-community interactions during vehicle stops and how to change those interactions for the better. Working closely with Dan Jurafsky's computational linguistics lab at Stanford, we developed an entirely new approach to examining and quantifying how vehicle stops unfold, unlocking the power of police body-worn cameras as a tool for change.[67]

Harnessing the potential of this technology in Oakland first required shifting the institutional norms and expectations around what body-worn camera footage is and what its purpose *could be*. Within many law enforcement agencies, the hundreds or thousands of hours of footage recorded daily by officers' body-worn cameras is thought of as evidence. This evidence is intended to shed light on a specific case should an investigation arise. Within the broader context of the criminal justice system, evidence tends to be used to exonerate or incriminate, and so many in law enforcement fear that routine collection of evidence about an agency's own practices could likewise be used to indict them. We worked to persuade the OPD executives that to make the most of the footage they were already collecting, it should be understood not as evidence, but as data. Data are neither "good" nor "bad," but can provide an inventory of the impact of an agency's practices

carried out in the aggregate by hundreds of officers per day, as opposed to solely being used to investigate allegations of misconduct against a particular officer. Body-worn camera footage can be leveraged to document and understand common patterns of engagement between the police and the public. Moreover, applying computational tools to the footage enables researchers, practitioners, and policymakers to, for the first time, measure and quantify police encounters to diagnose the health of police-community relations. Body-camera footage marries the strengths of big data with the dynamics of day-to-day interactions. These tools can be scaled to mine insights from any number of interactions while remaining sensitive enough to capture their subtle interpersonal qualities.

We worked as part of an interdisciplinary research team to develop novel computational tools to analyze body-camera footage and gain a better understanding of the mechanics and tenor of police-community interactions. In an initial study, we analyzed nearly one thousand vehicle stops in Oakland and found that officers consistently spoke less respectfully to Black drivers than to white drivers.[68] These racial disparities in police language remained even after controlling for the location of the stop, the outcome of the stop, the severity of the offense, and the race of the officer. We found, for example, that officers were more likely to offer reassurance ("No problem") to white drivers than to Black drivers and express concern for their well-being and safety ("Drive safe"). We found differences in the language officers used throughout the stop, even during the first seconds of the interaction, before the driver had much of a chance to speak. We also found disparities in officers' tone of voice, such that officers spoke in a more positive tone to white drivers than to Black drivers.[69] In fact, we found that when community members listened to clips from the body-camera footage of officers speaking in a more negative tone of voice, as officers are more likely to do with Black drivers, they rated having less institutional trust in the entire police department from which those clips originated and had a more negative view of police more generally.[70]

We also used our computational linguistics tools to map the conversational sequence and key events that take place as routine stops unfold over time, breaking them down and identifying particular stages.[71] The stages of a vehicle stop include: offering a greeting ("Hello, I'm Officer…"), providing a reason for the stop ("The reason I stopped you…") , asking for documents ("You have your driver's license, registration, and insurance?), asking for details ("Where do you live?"), a sanction ("The reason I'm citing you is for failure to yield to oncoming traffic"), and a closing ("All right. Drive safe"). This level of granularity enables researchers and police executives alike to explore how race may play a role at each turn of the interaction, and whether officers follow institutional norms and comply with relevant policies. For example, do officers consistently and clearly explain the reason they pulled the driver over? More than a style of communication, stating the reason for a stop amounts to providing a legal justification for the stop, which is

required by the Fourth Amendment.[72] The use of body-camera footage can thus aid efforts to ensure that police-community interactions are carried out in a constitutional and procedurally just manner.[73]

In addition to feeling like they are treated with respect and in a fair and transparent manner, community members also want assurance that they will be safe in interactions with police. Amid calls for police to de-escalate encounters with people from Black communities in particular, we are working to shed light on when and how routine interactions escalate. Using a different dataset, we are finding evidence of police escalation in the routine vehicle stops of Black male drivers. Vehicle stops that ultimately result in escalation differ in their conversational structure from the very beginning. In fact, the first forty-five words an officer utters – roughly the first twenty-seven seconds of a stop – predict whether that stop will end with the officer handcuffing, searching, or arresting the driver.[74] In these escalated stops, officers are more likely to issue commands as their opening words ("Keep your hands on the wheel"), and are less likely to tell drivers the reason they are being stopped ("The reason I stopped you is because your headlight is out"). When we asked Black men to listen to these escalated encounters in an online study, the clips evoked anxiety, suspicion, and worry that the officer would use force.[75]

After our initial paper was published and the findings about racial disparities in officers' language came out, Oakland community members told us that the research put numbers and data to their experiences. Feeling emboldened, they called on the department to do more to close the respect gap and address racial disparities. How did the OPD respond? We knew the respect gap in officers' language was certainly not in alignment with what those officers had been taught in the procedural-justice training program the entire department had gone through just a few years earlier. Training officers on procedural justice has been a popular policing tactic in recent years. It emphasizes fair treatment and transparent practices on the part of the officer when interacting with community members, regardless of the outcome of the encounter.[76] As part of the federal monitorship and in response to our stop-data report recommendations, the OPD was already planning to deploy a custom procedural-justice training as a follow-up to the standard training. Specifically, OPD leadership wanted to design their own agency-wide program to highlight concrete steps officers should take to put the principles of procedural justice into action in their local context.

Due to both the community's reaction and officers' own questions about our research on body-worn camera footage, OPD leadership came to us for help. We agreed to work with the agency's trainers to codevelop a training module on using respectful language during vehicle stops. This module provided the opportunity for sworn and civilian staff to openly discuss the research findings about the respect gap in language. The module also provided concrete, actionable, evidence-

based steps officers could take during vehicle stop interactions to improve the treatment of community members and decrease racial disparities.

When we codeveloped this module with the OPD, we were excited by the potential to demonstrate that body-worn camera footage could be used as a training tool, namely to give officers feedback on their behavior in the aggregate and to provide recommendations for how to improve based on their own agency's data. At the same time, however, we realized many of the officers would be resistant to this information for a variety of reasons. Just as after the release of our stop-data report, some officers felt attacked and wanted to discuss what they saw as the burning question on everyone's mind: "Do these findings mean that we are all racist?" Other officers felt like the data unfairly pointed the finger at them, blaming them for the agency's practices when they, unlike the command staff, have relatively little power within the organization. The findings gave rise to a host of identity-based threats, and so we set out to help the agency respond to them. Specifically, we leveraged the module to bolster internal procedural justice, which is fair treatment and transparent processes not just in police interactions with the external public but also internally, regarding policies and procedures within the organization.[77] We codesigned the module while keeping in mind several key social psychological strategies shown to mitigate bias and reduce threat, foster internal procedural justice, and support behavior changes.[78]

First, the module was delivered by OPD trainers who were well-liked and respected "insiders" and took place at the end of the training, once cohesion, comfort, and trust had been built among the group. This enabled an open and frank discussion about the research findings and their implications. Second, a main component of the module was a video Q&A between Eberhardt (the lead researcher) and then Deputy Chief LeRonne Armstrong (a high-ranking leader who was also well-liked and respected within the OPD) who asked hard questions on behalf of the officers, giving them "voice," which is a key tenet of procedural justice. Third, throughout the module, five concrete, actionable ways that officers could convey respect through their language (for example, by expressing concern for the driver's safety: "Take care tonight, sir") were highlighted. And finally, a brief role-play dialogue that the trainers performed grounded the information in a familiar and relevant scenario.

We did not stop there. We developed a method to compare the body-worn camera footage of police-community interactions before and after the training, to empirically examine training effectiveness. This was a bold move – as most police trainings designed to improve police-community relations are rarely evaluated. And if a training is evaluated at all, it tends to be by simply asking officers, "Did you like the training?"

Here, we focused on officers' use of communication techniques to convey respect to drivers. Compared to stops that occurred prior to the training, post-training

we found that officers employed more of these techniques. In particular, officers were more likely to express concern for drivers' safety, offer reassurance, and provide explicit reasons for why they made the stop.[79] More generally, examining footage pre- and post-training can help us determine whether and how the substance of what is taught in a training translates to specific behaviors that actually improve police-community interactions. These objective metrics can help a variety of stakeholders hold police departments accountable and improve upon the data used to do so.

For all the strides that have been made in Oakland, not all of the problems are solved. More than twenty years later, the OPD remains under federal oversight. There are limits to external methods of reforming policing. In fact, such methods can contribute to dichotomous evaluations of police departments as either broken or fixed.[80] But by taking a sociocultural approach to locating and combating racial bias, the focus shifts from whether a department, industry, or institution has managed to "fix" itself to whether they understand the ways the culture can contribute to where, how, and why racial disparities and bias manifest and spread.

By bringing together the psychological science of racial bias and culture and enacting a sociocultural approach for the purpose of reducing bias and racial disparities, we have provided an example of change that can be applied in the context of research-driven partnerships more broadly. Elsewhere, we have called this approach *intentional culture change*, and describe how to leverage the science of culture, bias, and inequality for behavioral, organizational, and societal change.[81] We provide a useful and actionable framework of seven core principles that can be applied to the issues of racial disparities in policing discussed here and to tackling social disparities across other domains.

Our research-driven collaboration with the OPD certainly looks different than what we as academics were originally trained to do. It has been difficult and time-consuming. It requires being out in the field, getting close to the social problems of the day, putting in the effort to learn practitioners' and stakeholders' worlds, cultivating meaningful relationships, identifying problems and being willing to work alongside key changemakers to fix them, all while navigating numerous cultural clashes and divides. Although it is certainly hard, we believe that science has a role to play, if not an obligation, to help society understand and reduce the racial disparities that can dramatically shape people's life outcomes.

Together, with stakeholders both within and outside the OPD, we constructed problem-focused research and explored change strategies across various levels of the organization's culture. This collaboration helped ensure that the strategies we developed were feasible, practical, and tailored to the context, and therefore had a greater likelihood of being implemented with fidelity and of being effective. This

type of work involves being humble and curious, listening more than talking, and not being discouraged by the messiness and complexity of the real world. By "getting proximate,"[82] researchers can learn more about context, establish a presence in the organization – hearing from those embedded within and outside it – and learn more about its intricacies, needs, challenges, and unique levers for change that might be available, or that could be created anew.

In the summer of 2021, twenty-one years after they filed the class-action lawsuit against the City of Oakland, civil rights attorneys John Burris and Jim Chanin wrote in a brief filed with the federal court: "The Oakland Police Department has moved from being one of the worst police departments in the San Francisco Bay Area to being one of the best police departments in comparable cities in the country."[83] Indeed, in our most recent conversations with Burris and Chanin, they describe a changed department. No longer are there illegal detentions of Oakland's Black residents. No longer are officers arresting people for "resisting arrest" without any other underlying offenses, which, according to Burris, "was pretty common back in the day." He says that the cases that "have a dramatic impact on people's lives are not happening at the same high level as before."

So much has changed. Most notably, neither Burris nor Chanin have the "beat up" cases anymore. Years ago, it was common for Black Oakland residents to show up at their law offices bruised and battered, claiming they had been assaulted by Oakland police officers, that they had been publicly humiliated, their lives undone. "It's been almost twenty years since these clients have come in," Burris says, "and some of these people were beaten as badly as Rodney King – they had these cases that generated outrage." Without a doubt, both Burris and Chanin believe that the negotiated settlement agreement they entered into with the City of Oakland on behalf of people who suffered such horrific harms at the hands of Oakland police officers had a "dramatic impact on the type of brutality officers engage in." Those brutality cases have all but disappeared. "The culture has changed there [at the OPD]," says Chanin. The number of stops of innocent Black residents has dropped dramatically, and when these residents are stopped, they are treated with more dignity and respect. It was a collective effort, for which we can all feel proud.

But change can be fragile. This is what worries Burris and Chanin most. It worries us too. There is no doubt that the OPD made dramatic changes – when crime was declining. But now crime is rising again. From 2022 to 2023, for example, motor vehicle theft alone increased by 36 percent.[84] Oakland's violent crime rates are significantly higher than other cities in California, and nearly two and a half times as high as in San Francisco, just a short ride across the Bay Bridge.[85] It is difficult to turn on the local news without hearing about the troubles in Oakland. And the numbers are no more comforting when comparing 2022 with a five-year average: motor vehicle theft is up by 21 percent, homicide is up by 23 percent, carjacking up by 53 percent, commercial burglary up by 56 percent.[86] The conditions are moving closer to

those that coincided with the formation of the Riders. In this moment of crisis, Oakland residents are desperately seeking solutions. Something has to be done.

Burris and Chanin have been on the job for more than twenty years now – protecting Oakland residents from those who have been sworn to protect them. But their stint is drawing to a close. What will happen when they leave? What will happen when the federal monitor leaves – when a federal judge is no longer presiding over the case? What will happen when *we* leave? Both Burris and Chanin are "hopeful, but cautious." Such is the work of change.

ABOUT THE AUTHORS

Rebecca C. Hetey is Associate Director of Criminal Justice Partnerships and Research Scientist at Stanford SPARQ at Stanford University. She studies race and the criminal justice system and works with practitioners to develop strategies to reduce racial disparities and improve police-community relations. She has published in such journals as *Proceedings of the National Academy of Sciences*, *Psychological Science*, and *Current Directions in Psychological Science*.

MarYam G. Hamedani is Executive Director and Senior Research Scientist at Stanford SPARQ at Stanford University. At SPARQ, she creates opportunities for researchers and practitioners to learn from one another in mutually beneficial partnerships. Her expertise is in harnessing the power of culture to support organizational and societal change and disrupting cultural defaults that lead to bias and inequality. She has published in such journals as *American Psychologist*, *Psychological Science*, and *Journal of Personality and Social Psychology*.

Hazel Rose Markus, a Fellow of the American Academy since 1994, is Davis-Brack Professor in the Behavioral Sciences, Professor of Psychology, and Cofounder and Codirector of Stanford SPARQ at Stanford University. She is the author of four books, including *Doing Race: 21 Essays for the 21st Century* (with Paula M. L. Moya, 2010) and *Clash! How to Thrive in a Multicultural World* (with Alana Conner, 2014), and editor of *Facing Social Class: How Societal Rank Influences Interaction* (with Susan T. Fiske, 2012) and *Social Psychology* (with Saul Kassin and Steven Fein, 2024). She is also a member of the National Academy of Sciences.

Jennifer L. Eberhardt, a Fellow of the American Academy since 2016, is William R. Kimball Professor at the Stanford Graduate School of Business, Professor of Psychology, and Cofounder and Codirector of Stanford SPARQ at Stanford University. Stanford SPARQ is a behavioral science "do tank" that builds research-driven partnerships with industry leaders and changemakers to combat bias, reduce disparities, and drive culture change. She is the author of *Biased: Uncovering the Hidden Prejudice That Shapes What We See, Think, and Do* (2019) and a recipient of the MacArthur "genius grant" Fellowship (2014). She is also a member of the National Academy of Sciences and the American Philosophical Society.

ENDNOTES

[1] Jennifer L. Eberhardt, *Biased: Uncovering the Hidden Prejudice That Shapes What We See, Think, and Do* (New York: Viking, 2019).

[2] Ali Winston and Darwin BondGraham, *The Riders Come Out at Night: Brutality, Corruption, and Cover-Up in Oakland* (New York: Atria Books, 2023).

[3] Kevin Fagan and Henry K. Lee, "Oakland Reins in Its Rough 'Riders' / In a Tough Neighborhood, Four Cops Pushed the Limits while Ruling Their Beat with an Iron Fist," *SFGate*, October 2, 2000, https://www.sfgate.com/bayarea/article/Oakland-Reins-In-Its-Rough-Riders-In-a-tough-3237415.php.

[4] Winston and BondGraham, *The Riders Come Out at Night*.

[5] Eberhardt, *Biased*, 76–77.

[6] Winston and BondGraham, *The Riders Come Out at Night*.

[7] Monica Anderson, "Vast Majority of Blacks View the Criminal Justice System as Unfair," *Pew Research Center*, August 12, 2014, https://www.pewresearch.org/short-reads/2014/08/12/vast-majority-of-blacks-view-the-criminal-justice-system-as-unfair; David Brooks, "The Culture of Policing is Broken," *The Atlantic*, June 16, 2020, https://www.theatlantic.com/ideas/archive/2020/06/how-police-brutality-gets-made/613030; Civil Rights Division, *Investigation of the Ferguson Police Department* (Washington, D.C.: U.S. Department of Justice, 2015); David Cole, *No Equal Justice: Race and Class in the American Criminal Justice System* (New York: The New Press, 1999); Sean Illing, "Why the Policing Problem Isn't about 'a Few Bad Apples,'" *Vox*, June 6, 2020, https://www.vox.com/identities/2020/6/2/21276799/george-floyd-protest-criminal-justice-paul-butler; Justin McCarthy, "Americans Remain Steadfast on Policing Reform Needs in 2022," Gallup, May 27, 2022, https://news.gallup.com/poll/393119/americans-remain-steadfast-policing-reform-needs-2022.aspx; President's Task Force on 21st Century Policing, *Final Report of the President's Task Force on 21st Century Policing* (Washington, D.C.: Office of Community Oriented Policing Services, 2015); Sue Rahr, "The Myth Propelling America's Violent Police Culture," *The Atlantic*, January 31, 2023, https://www.theatlantic.com/ideas/archive/2023/01/police-brutality-shootings-derek-chauvin/672873; Jillian K. Swencionis and Phillip Atiba Goff, "The Psychological Science of Racial Bias and Policing," *Psychology, Public Policy, and Law* 23 (4) (2017): 398–409; and Sophie Trawalter, D-J Bart-Plange, and Kelly M. Hoffman, "A Socioecological Psychology of Racism: Making Structures and History More Visible," *Current Opinion in Psychology* 32 (2020): 47–51.

[8] John F. Dovidio and Samuel L. Gaertner, "On the Nature of Contemporary Prejudice: The Causes, Consequences, and Challenges of Aversive Racism," in *Confronting Racism: The Problem and the Response*, ed. Jennifer L. Eberhardt and Susan T. Fiske (Thousand Oaks, Calif.: Sage, 1998); Eberhardt, *Biased;* James M. Jones, *Prejudice and Racism* (New York: McGraw-Hill, 1997); Hazel Rose Markus and Paula M. L. Moya, *Doing Race: 21 Essays for the 21st Century* (New York: W. W. Norton & Company, 2010); Jennifer A. Richeson and J. Nicole Shelton, "When Prejudice Does Not Pay: Effects of Interracial Contact on Executive Function," *Psychological Science* 14 (3) (2003): 287–290; Phia S. Salter, Glenn Adams, and Michael J. Perez, "Racism in the Structure of Everyday Worlds: A Cultural-Psychological Perspective," *Current Directions in Psychological Science* 27 (3) (2018): 150–155; and Sophie Trawalter, Gerald D. Higginbotham, and Kyshia Henderson, "Social Psychological Research on Racism and the Importance of Historical Context: Implications for Policy," *Current Directions in Psychological Science* 31 (6) (2022): 493–499.

[9] Alan P. Fiske, Shinobu Kitayama, Hazel Rose Markus, and Richard E. Nisbett, "The Cultural Matrix of Social Psychology," in *The Handbook of Social Psychology*, ed. Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey (New York: McGraw Hill, 1998), 915–981; Steven J. Heine, *Cultural Psychology* (New York: W. W. Norton & Company, 2020); Yoshihisa Kashima, "What is Culture For?" in *The Handbook of Culture and Psychology*, ed. David Matsumoto and Hyisung C. Hwang (New York: Oxford University Press, 2019), 123–160; Hazel Rose Markus and Shinobu Kitayama, "Cultures and Selves: A Cycle of Mutual Constitution," *Perspectives on Psychological Science* 5 (4) (2010): 420–430; and Michael W. Morris, Ying-yi Hong, Chi-yue Chiu, and Zhi Liu, "Normology: Integrating Insights about Social Norms to Understand Cultural Dynamics," *Organizational Behavior and Human Decision Processes* 129 (2015): 1–13.

[10] Glenn Adams, Monica Biernat, Nyla R. Branscombe, et al., "Beyond Prejudice: Toward a Sociocultural Psychology of Racism and Oppression," in *Commemorating Brown: The Social Psychology of Racism and Discrimination*, ed. Glenn Adams, Monica Biernat, Nyla R. Branscombe, Christian S. Crandall, and Lawrence S. Wrightsman (Washington, D.C.: American Psychological Association, 2008), 215–246; Jones, *Prejudice and Racism;* Markus and Moya, *Doing Race: 21 Essays for the 21st Century*; B. Keith Payne, Heidi A. Vuletich, and Jazmin L. Brown-Iannuzzi, "Historical Roots of Implicit Bias in Slavery," *Proceedings of the National Academy of Sciences* 116 (24) (2019): 11693–11698; and Salter, Adams, and Perez, "Racism in the Structure of Everyday Worlds: A Cultural-Psychological Perspective." In this volume, see Manuel J. Galvan and B. Keith Payne, "Implicit Bias as a Cognitive Manifestation of Systemic Racism," *Dædalus* 153 (1) (Winter 2024): 106–122, https://www.amacad.org/publication/implicit-bias-cognitive-manifestation-systemic-racism.

[11] Fiske, Kitayama, Markus, and Nisbett, "The Cultural Matrix of Social Psychology"; Hazel Rose Markus and Alana Conner, *Clash!: How to Thrive in a Multicultural World* (New York: Plume, 2014); and Markus and Kitayama, "Cultures and Selves: A Cycle of Mutual Constitution."

[12] MarYam G. Hamedani, Hazel Rose Markus, Rebecca C. Hetey, and Jennifer L. Eberhardt, "We Built This Culture (So We Can Change It): Seven Principles for Intentional Culture Change," *American Psychologist* (advance online publication, 2023), https://doi.org/10.1037/amp0001209; Sapna Cheryan and Hazel Rose Markus, "Masculine Defaults: Identifying and Mitigating Hidden Cultural Biases," *Psychological Review* 127 (6) (2020): 1022–1052; MarYam G. Hamedani and Hazel Rose Markus, "Understanding Culture Clashes and Catalyzing Change: A Culture Cycle Approach," *Frontiers in Psychology* 10 (2019); and Cayce J. Hook and Hazel Rose Markus, "Health in the United States: Are Appeals to Choice and Personal Responsibility Making Americans Sick?" *Perspectives on Psychological Science* 15 (3) (2020): 643–664.

[13] Hamedani, Markus, Hetey, and Eberhardt, "We Built This Culture (So We Can Change It): Seven Principles for Intentional Culture Change"; Hook and Markus, "Health in the United States: Are Appeals to Choice and Personal Responsibility Making Americans Sick?"; Sarah Lyons-Padilla, Hazel Rose Markus, Ashby Monk, et al., "Race Influences Professional Investors' Financial Judgments," *Proceedings of the National Academy of Sciences* 116 (35) (2019): 17225–17230; Nicole M. Stephens, MarYam G. Hamedani, and Sarah S. M. Townsend, "Difference Matters: Teaching Students a Contextual Theory of Difference Can Help Them Succeed," *Perspectives on Psychological Science* 14 (2) (2019): 156–174; Catherine C. Thomas, Nicholas G. Otis, Justin R. Abraham, et al., "Toward a Science of Delivering Aid with Dignity: Experimental Evidence and Local Forecasts from Kenya," *Proceedings of the National Academy of Sciences* 117 (27) (2020): 15546–15553;

and Catherine C. Thomas, Gregory M. Walton, Ellen C. Reinhart, and Hazel Rose Markus, "Mitigating Welfare-Related Prejudice and Partisanship among U.S. Conservatives with Moral Reframing of a Universal Basic Income Policy," *Journal of Experimental Social Psychology* 105 (2023).

[14] Nicholas P. Camp, Rob Voigt, Dan Jurafsky, and Jennifer L. Eberhardt, "The Thin Blue Waveform: Racial Disparities in Officer Prosody Undermine Institutional Trust in the Police," *Journal of Personality and Social Psychology* 121 (6) (2021): 1157–1171; Nicholas P. Camp, Rob Voigt, MarYam G. Hamedani, et al., "Leveraging Body-Worn Camera Footage to Assess the Effects of Training on Officer Communication During Traffic Stops" (unpublished manuscript, 2023); Eberhardt, *Biased*; Anjalie Field, Prateek Verma, Nay San, et al., "Developing Speech Processing Pipelines for Police Accountability," *INTERSPEECH 2023*, https://doi.org/10.21437/Interspeech.2023-2109; Rebecca C. Hetey, "Implicit Bias, the Power of Institutions, and How to Reduce Racial Disparities in Policing," in *Bias in the Law: A Definitive Look at Racial Prejudice in the U.S. Criminal Justice System*, ed. Joseph Avery and Joel Cooper (Lanham, Md.: Lexington Books, 2020), 37–66; Rebecca C. Hetey, Benoît Monin, Amrita Maitreyi, and Jennifer L. Eberhardt, *Data for Change: A Statistical Analysis of Police Stops, Searches, Handcuffings, and Arrests in Oakland, Calif., 2013 – 2014* (Stanford, Calif.: Stanford SPARQ, 2016); Vinodkumar Prabhakaran, Camilla Griffiths, Hang Su, et al., "Detecting Institutional Dialog Acts in Police Traffic Stops," *Transactions of the Association for Computational Linguistics* 6 (2018): 467–481; Eugenia H. Rho, Maggie Harrington, Yuyang Zhong, et al., "Escalated Police Stops of Black Men Are Linguistically and Psychologically Distinct in Their Earliest Moments," *Proceedings of the National Academy of Sciences* 120 (23) (2023): e2216162120; and Rob Voigt, Nicholas P. Camp, Vinodkumar Prabhakaran, et al., "Language from Police Body Camera Footage Shows Racial Disparities in Officer Respect," *Proceedings of the National Academy of Sciences* 114 (25) (2017): 6521–6526.

[15] Christine Eith and Matthew R. Durose, *Contacts Between Police and the Public, 2008* (Washington, D.C.: Office of Justice Programs, Bureau of Justice Statistics, 2011).

[16] Erika Harrell and Elizabeth Davis, *Contacts Between Police and the Public, 2018 – Statistical Tables* (Washington, D.C.: Office of Justice Programs, Bureau of Justice Statistics, 2020).

[17] Ibid.; National Institute of Justice, *Racial Profiling and Traffic Stops* (Washington, D.C.: National Institute of Justice, 2013); Ian Ayres and Jonathan Borowsky, *A Study of Racially Disparate Outcomes in the Los Angeles Police Department* (Los Angeles: ACLU of Southern California, 2008); and William R. Smith, Donald Tomaskovic-Devey, Matthew T. Zingraff, et al., *The North Carolina Highway Traffic Study* (Washington, D.C.: National Institute of Justice, 2003).

[18] Frank R. Baumgartner, Derek A. Epp, and Kelsey Shoub, *Suspect Citizens: What 20 Million Traffic Stops Tell Us About Policing and Race* (Cambridge: Cambridge University Press, 2018).

[19] Sarah Ravani, Alejandro Serrano, and Steve Rubenstein, "Oakland Police Have Faced Scandals and Controversies for Two Decades. Here's a Look Back," *San Francisco Chronicle*, February 15, 2023, https://www.sfchronicle.com/projects/2023/oakland-police-chief-timeline.

[20] Winston and BondGraham, *The Riders Come Out at Night*.

[21] Larry D. Hatfield and Janine DeFao, "Ex-Cop Who Worked With 'The Riders' Sues Oakland," *SFGate*, February 1, 2002, https://www.sfgate.com/news/article/ex-cop-who-worked-with-the-riders-sues-oakland-2877875.php.

[22] Winston and BondGraham, *The Riders Come Out at Night*, 83.

[23] *Delphine Allen et al. v. City of Oakland*, C00-4599 (N.D. Calif. 2000).

[24] Eberhardt, *Biased*.

[25] Nicholas P. Camp, "Institutional Interactions and Racial Inequality in Policing: How Everyday Encounters Bridge Individuals, Organizations, and Institutions," *Social and Personality Psychology Compass* (advance online publication, 2023): e12930, https://doi.org/10.1111/spc3.12930.

[26] Camp, Voigt, Jurafsky, and Eberhardt, "The Thin Blue Waveform: Racial Disparities in Officer Prosody Undermine Institutional Trust in the Police."

[27] Michele J. Gelfand, Jana L. Raver, Lisa Nishii, et al., "Differences Between Tight and Loose Cultures: A 33-Nation Study," *Science* 332 (6033) (2011): 1100–1104.

[28] Eugene A. Paoline III, "Taking Stock: Toward a Richer Understanding of Police Culture," *Journal of Criminal Justice* 31 (3) (2003): 199–214.

[29] For explanations of the culture cycle, see Fiske, Kitayama, Markus, and Nisbett, "The Cultural Matrix of Social Psychology"; Markus and Kitayama, "Cultures and Selves: A Cycle of Mutual Constitution"; and Markus and Conner, *Clash!* For examples of how to apply the culture cycle as a tool for culture change in society and within institutions and organizations, see Hamedani, Markus, Hetey, and Eberhardt, "We Built This Culture (So We Can Change It): Seven Principles for Intentional Culture Change"; Cheryan and Markus, "Masculine Defaults," *Psychological Review* 127 (6) (2020): 1022–1052; Hamedani and Markus, "Understanding Culture Clashes and Catalyzing Change"; and Hook and Markus, "Health in the United States: Are Appeals to Choice and Personal Responsibility Making Americans Sick?"

There are a few notes that we would like to provide to accompany the culture cycle. First, cultures are always dynamic systems. Second, the culture cycle includes and incorporates organizational and institutional structures and dynamics. The concepts of "culture" (such as collective beliefs, practices, and products) and "structure" (such as societal institutions and organizations) are integrated rather than separated. Third, culture cycles are embedded in broader historical, ecological, and evolutionary systems that interact with and exert influence on a given culture, both in the past and present. Fourth, different cultures can also interact with and influence one another, sometimes in expected and sometimes in unexpected ways. And fifth, while there are various other models that represent culture as a multilevel system that have various aims and distinctions, they often share the goal of delineating the key features of culture in a simplified, usable form.

[30] Eric Shuman, Siwar Hasan-Aslih, Martijn van Zomeren, et al., "Protest Movements Involving Limited Violence Can Sometimes Be Effective: Evidence from the 2020 BlackLivesMatter Protests," *Proceedings of the National Academy of Sciences* 119 (14) (2022): e2118990119.

[31] Paoline, "Taking Stock: Toward a Richer Understanding of Police Culture"; and John Van Maanen, "Working the Street: A Developmental View of Police Behavior," in *The Potential for Reform of Criminal Justice*, ed. Herbert Jacob (Beverly Hills: Sage Publications, 1974), 83–130.

[32] Warren Friedman, Arthur J. Lurigio, Richard Greenleaf, and Stephanie Albertson, "Encounters Between Police Officers and Youths: The Social Costs of Disrespect," *Journal of Crime and Justice* 27 (2) (2004): 1–25; Lorraine Mazerolle, Emma Antrobus, Sarah Bennett, and Tom R. Tyler, "Shaping Citizen Perceptions of Police Legitimacy: A Randomized Field Trial of Procedural Justice," *Criminology* 51 (1) (2013): 33–63; Wesley G. Skogan, Maarten Van Craen, and Cari Hennessy, "Training Police for Procedural Justice," *Journal of Experimental Criminology* 11 (2015): 319–334; and Tom R. Tyler and Cheryl J. Wakslak, "Profiling and Police Legitimacy: Procedural Justice, Attributions of Motive, and Acceptance of Police Authority," *Criminology* 42 (2) (2004): 253–282.

[33] Eric Bailey, "'He's Made Them Proud of Oakland,'" *Los Angeles Times*, February 10, 2000, https://www.latimes.com/archives/la-xpm-2000-feb-10-mn-62923-story.html; Fagan and Lee, "Oakland Reins in Its Rough 'Riders'"; and Doug Foster, "Jerry Brown: The Outsider," *Rolling Stone*, April 15, 1999, https://www.rollingstone.com/politics/politics-news/jerry-brown-the-outsider-180047. For the mayor's "campaign against crime and grime," see Winston and BondGraham, *The Riders Come Out at Night*, 38.

[34] KRON-TV Assignment Four, "The People and the Police: Oakland (1974)," https://www.youtube.com/watch?v=B-lhlB5r7Uw.

[35] Ben Brucato, "Policing Race and Racing Police: The Origin of U.S. Police in Slave Patrols," *Social Justice* 47 (3–4) (2020): 115–136; Philip L. Reichel, "Southern Slave Patrols as a Transitional Police Type," *American Journal of Police* 7 (2) (1988): 51–78; Larry H. Spruill, "Slave Patrols, 'Packs of Negro Dogs' and Policing Black Communities," *Phylon* 53 (1) (2016): 42–66; and K. B. Turner, David Giacopassi, and Margaret Vandiver, "Ignoring the Past: Coverage of Slavery and Slave Patrols in Criminal Justice Texts," *Journal of Criminal Justice Education* 17 (1) (2006): 181–195.

[36] Hetey, Monin, Maitreyi, and Eberhardt, *Data for Change*.

[37] Frank R. Baumgartner, Leah Christiani, Derek A. Epp, et al., "Racial Disparities in Traffic Stop Outcomes," *Duke Forum for Law & Social Change* 9 (1) (2017): 21–54; Andrew Gelman, Jeffrey Fagan, and Alex Kiss, "An Analysis of the New York City Police Department's 'Stop-and-Frisk' Policy in the Context of Claims of Racial Bias," *Journal of the American Statistical Association* 102 (479) (2007): 813–823; Sharad Goel, Justin M. Rao, and Ravi Shroff, "Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy," *The Annals of Applied Statistics* 10 (1) (2016): 365–394; Emma Pierson, Camelia Simoiu, Jan Overgoor, et al., "A Large-Scale Analysis of Racial Disparities in Police Stops across the United States," *Nature Human Behaviour* 4 (7) (2020): 736–745; and Civil Rights Division, *Investigation of the Baltimore City Police Department* (Washington, D.C.: U.S. Department of Justice, 2016).

[38] Eberhardt, *Biased;* Jones, *Prejudice and Racism;* Markus and Moya, *Doing Race: 21 Essays for the 21st Century*; Salter, Adams, and Perez, "Racism in the Structure of Everyday Worlds: A Cultural-Psychological Perspective"; and Claude M. Steele, *Whistling Vivaldi: How Stereotypes Affect Us and What We Can Do* (New York: W.W. Norton & Company, 2011).

[39] Hetey, Monin, Maitreyi, and Eberhardt, *Data for Change*.

[40] Pierson, Simoiu, Overgoor, et al., "A Large-Scale Analysis."

[41] Karen E. Fields and Barbara J. Fields, *Racecraft: The Soul of Inequality in American Life* (New York: Verso, 2012); James M. Jones and John F. Dovidio, "Change, Challenge, and Prospects for a Diversity Paradigm in Social Psychology," *Social Issues and Policy Review* 12 (1) (2018): 7–56; Eduardo Bonilla-Silva, "What Makes Systemic Racism Systemic?"

Sociological Inquiry 91 (3) (2021): 513–533; Michael Omi and Howard Winant, *Racial Formation in the United States* (London: Routledge, 2014); and Joe R. Feagin, *Racist America: Roots, Current Realities, and Future Reparations* (New York: Routledge, 2000).

[42] Eberhardt, *Biased;* and Steele, *Whistling Vivaldi*.

[43] Jennifer L. Eberhardt, Phillip Atiba Goff, Valerie J. Purdie, and Paul G. Davies, "Seeing Black: Race, Crime, and Visual Processing," *Journal of Personality and Social Psychology* 87 (6) (2004): 876–893; and Charles R. Epp, Steven Maynard-Moody, and Donald Haider-Markel, *Pulled Over: How Police Stops Define Race and Citizenship* (Chicago: The University of Chicago Press, 2014).

[44] Joseph Avery and Joel Cooper, *Bias in the Law: A Definitive Look at Racial Prejudice in the U.S. Criminal Justice System* (Lanham, Md.: Lexington Books, 2020); Joshua Correll, Bernadette Park, Charles M. Judd, and Bernd Wittenbrink, "The Police Officer's Dilemma: Using Ethnicity to Disambiguate Potentially Threatening Individuals," *Journal of Personality and Social Psychology* 83 (6) (2002): 1314–1329; Joshua Correll, Bernadette Park, Charles M. Judd, et al., "Across the Thin Blue Line: Police Officers and Racial Bias in the Decision to Shoot," *Journal of Personality and Social Psychology* 92 (6) (2007): 1006–1023; Patricia G. Devine, "Stereotypes and Prejudice: Their Automatic and Controlled Components," *Journal of Personality and Social Psychology* 56 (1) (1989): 5–18; Patricia G. Devine and Andrew J. Elliot, "Are Racial Stereotypes Really Fading? The Princeton Trilogy Revisited," *Personality and Social Psychology Bulletin* 21 (11) (1995): 1139–1150; Eberhardt, *Biased*; Jennifer L. Eberhardt, "Enduring Racial Associations: African Americans, Crime, and Animal Imagery," in *Doing Race: 21 Essays for the 21st Century*, ed. Markus and Moya, 439–457; Jennifer L. Eberhardt, Paul G. Davies, Valerie J. Purdie-Vaughns, and Sheri Lynn Johnson, "Looking Deathworthy: Perceived Stereotypicality of Black Defendants Predicts Capital-Sentencing Outcomes," *Psychological Science* 17 (5) (2006): 383–386; Epp, Maynard-Moody, and Haider-Markel, *Pulled Over;* Jack Glaser, *Suspect Race: Causes and Consequences of Racial Profiling* (New York: Oxford University Press, 2015); Alison V. Hall, Erika V. Hall, and Jamie L. Perry, "Black and Blue: Exploring Racial Bias and Law Enforcement in the Killings of Unarmed Black Male Civilians," *American Psychologist* 71 (3) (2016): 175–186; Hetey, "Implicit Bias, the Power of Institutions, and How to Reduce Racial Disparities in Policing"; Julian M. Rucker and Jennifer A. Richeson, "Toward an Understanding of Structural Racism: Implications for Criminal Justice," *Science* 374 (6565) (2021): 286–290; and Samuel R. Sommers and Phoebe C. Ellsworth, "White Juror Bias: An Investigation of Prejudice Against Black Defendants in the American Courtroom," *Psychology, Public Policy, and Law* 7 (1) (2001): 201–229.

[45] The concept of crime itself was racialized at the turn of the twentieth century and attached to Black people as an entire group–but not white people, for whom crime is seen as an individual failing–as a way to cement ideas of Black inferiority and uphold white supremacy. See Khalil Gibran Muhammed, *The Condemnation of Blackness: Race, Crime, and the Making of Modern Urban America* (Cambridge, Mass.: Harvard University Press, 2011); and Ben Grunwald, Julian Nyarko, and John Rappaport, "Police Agencies on Facebook Overreport on Black Suspects," *Proceedings of the National Academy of Sciences* 119 (45) (2022): e2203089119.

[46] Eberhardt, Goff, Purdie, and Davies, "Seeing Black"; and Phillip Atiba Goff, Jennifer L. Eberhardt, Melissa J. Williams, and Matthew Christian Jackson, "Not Yet Human: Implicit Knowledge, Historical Dehumanization, and Contemporary Consequences," *Journal of Personality and Social Psychology* 94 (2) (2008): 292–306.

[47] Eberhardt, Goff, Purdie, and Davies, "Seeing Black."

[48] Eberhardt, *Biased*.

[49] Robert B. Cialdini, Raymond R. Reno, and Carl A. Kallgren, "A Focus Theory of Normative Conduct: Recycling the Concept of Norms to Reduce Littering in Public Places," *Journal of Personality and Social Psychology* 58 (6) (1990): 1015–1026; Noah J. Goldstein, Robert B. Cialdini, and Vladas Griskevicius, "A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels," *Journal of Consumer Research* 35 (3) (2008): 472–482; Deborah A. Prentice and Dale T. Miller, "Pluralistic Ignorance and Alcohol Use on Campus: Some Consequences of Misperceiving the Social Norm," *Journal of Personality and Social Psychology* 64 (2) (1993): 243–256; and P. Wesley Schultz, Jessica M. Nolan, Robert B. Cialdini, et al., "The Constructive, Destructive, and Reconstructive Power of Social Norms," *Psychological Science* 18 (5) (2007): 429–434.

[50] Aaron C. Kay, Danielle Gaucher, Jennifer M. Peach, et al., "Inequality, Discrimination, and the Power of the Status Quo: Direct Evidence for a Motivation to See the Way Things Are as the Way They Should Be," *Journal of Personality and Social Psychology* 97 (3) (2009): 421–434.

[51] Epp, Maynard-Moody, and Haider-Markel, *Pulled Over*.

[52] *Whren v. United States*, 95-5841, 517 (D.C. Cir. 1996).

[53] David A. Harris, "'Driving While Black' and All Other Traffic Offenses: The Supreme Court and Pretextual Traffic Stops," *The Journal of Criminal Law and Criminology* 87 (2) (1997): 544–582.

[54] Brian J. O'Donnell, "*Whren v. United States*: An Abrupt End to the Debate Over Pretextual Stops," *Maine Law Review* 49 (1) (1997): 207–234.

[55] Bradley R. Haywood, "Ending Race-Based Pretextual Stops: Strategies for Eliminating America's Most Egregious Police Practice," *Richmond Public Interest Law Review* 26 (1) (2022): 47–84.

[56] Epp, Maynard-Moody, and Haider-Markel, *Pulled Over*.

[57] Justin D. Levinson, Robert J. Smith, and Koichi Hioki, "Race and Retribution: An Empirical Study of Implicit Bias and Punishment in America," *UC Davis Law Review* 53 (2) (2019): 839–892.

[58] Frances E. Aboud, "The Formation of In-Group Favoritism and Out-Group Prejudice in Young Children: Are They Distinct Attitudes?" *Developmental Psychology* 39 (1) (2003): 48–60; Marilynn B. Brewer, "The Psychology of Prejudice: Ingroup Love and Outgroup Hate?" *Journal of Social Issues* 55 (3) (1999): 429–444; Mina Cikara, Emile G. Bruneau, and Rebecca R. Saxe, "Us and Them: Intergroup Failures of Empathy," *Current Directions in Psychological Science* 20 (3) (2011): 149–153; Anthony G. Greenwald and Thomas F. Pettigrew, "With Malice toward None and Charity for Some: Ingroup Favoritism Enables Discrimination," *American Psychologist* 69 (7) (2014): 669–684; and Shana Levin and Jim Sidanius, "Social Dominance and Social Identity in the United States and Israel: Ingroup Favoritism or Outgroup Derogation?" *Political Psychology* 20 (1) (1999): 99–126.

[59] Rebecca C. Hetey and Jennifer L. Eberhardt, "Racial Disparities in Incarceration Increase Acceptance of Punitive Policies," *Psychological Science* 25 (10) (2014): 1949–1954.

[60] Bill Hutchinson, "From 'BBQ Becky' to 'Golfcart Gail,' List of Unnecessary 911 Calls Made on Blacks Continues to Grow," *ABC News*, October 19, 2018, https://abcnews

.go.com/US/bbq-becky-golfcart-gail-list-unnecessary-911-calls/story?id=58584961; Racial and Identity Profiling Advisory Board, *Annual Report 2019* (Sacramento: California Department of Justice, 2019); Racial and Identity Profiling Advisory Board, *Annual Report 2020* (Sacramento: California Department of Justice, 2020); and Lisa Thurau and Bob Stewart, "Avoiding 'Profiling by Proxy,'" Vera Institute of Justice, March 13, 2015, https://www.vera.org/news/police-perspectives/avoiding-profiling-by-proxy.

61 Rich Morin and Renee Stepler, *The Racial Confidence Gap in Police Performance* (Washington, D.C.: Pew Research Center, 2016).

62 Epp, Maynard-Moody, and Haider-Markel, *Pulled Over*; President's Task Force on 21st Century Policing, *Final Report of the President's Task Force on 21st Century Policing*; Jason Sunshine and Tom R. Tyler, "The Role of Procedural Justice and Legitimacy in Shaping Public Support for Policing," *Law & Society Review* 37 (3) (2003): 513–548; Tyler and Wakslak, "Profiling and Police Legitimacy: Procedural Justice, Attributions of Motive, and Acceptance of Police Authority"; and Tom R. Tyler, Phillip Atiba Goff, and Robert J. MacCoun, "The Impact of Psychological Science on Policing in the United States: Procedural Justice, Legitimacy, and Effective Law Enforcement," *Psychological Science in the Public Interest* 16 (3) (2015): 75–109.

63 Eberhardt, *Biased*.

64 Oakland Police Department, *2016 – 2018 Racial Impact Report* (Oakland: Office of Chief of Police, 2019).

65 Adam Grant, *Think Again: The Power of Knowing What You Don't Know* (New York: Viking, 2021).

66 Jennifer L. Eberhardt, "Fifty Recommendations to Mitigate Racial Disparities and Improve Police-Community Relations," in *Strategies for Change: Research Initiatives and Recommendations to Improve Police-Community Relations in Oakland, Calif.* (Stanford, Calif.: Stanford SPARQ, 2016), 40–57.

67 Camp, Voigt, Jurafsky, and Eberhardt, "The Thin Blue Waveform: Racial Disparities in Officer Prosody Undermine Institutional Trust in the Police"; Field, Verma, San, et al., "Developing Speech Processing Pipelines for Police Accountability"; Prabhakaran, Griffiths, Su, et al., "Detecting Institutional Dialog Acts in Police Traffic Stops"; Rho, Harrington, Zhong, et al., "Escalated Police Stops of Black Men Are Linguistically and Psychologically Distinct in Their Earliest Moments"; and Voigt, Camp, Prabhakaran, et al., "Language from Police Body Camera Footage Shows Racial Disparities in Officer Respect."

68 Voigt, Camp, Prabhakaran, et al., "Language from Police Body Camera Footage Shows Racial Disparities in Officer Respect."

69 Camp, Voigt, Jurafsky, and Eberhardt, "The Thin Blue Waveform: Racial Disparities in Officer Prosody Undermine Institutional Trust in the Police."

70 Ibid.

71 Prabhakaran, Griffiths, Su, et al., "Detecting Institutional Dialog Acts in Police Traffic Stops."

72 *Brendlin v. California*, 06–8120, 551 (S.C. Calif. 2007).

73 Tom R. Tyler, "Procedural Justice, Legitimacy, and the Effective Rule of Law," *Crime and Justice* 30 (2003): 283–357.

[74] Rho, Harrington, Zhong, et al., "Escalated Police Stops of Black Men Are Linguistically and Psychologically Distinct in Their Earliest Moments."

[75] Ibid.

[76] President's Task Force on 21st Century Policing, *Final Report of the President's Task Force on 21st Century Policing*; and Tom R. Tyler and Yuen J. Huo, *Trust in the Law: Encouraging Public Cooperation with the Police and Courts* (New York: Russell Sage Foundation, 2002).

[77] James D. Carr and Sheila Royo Maxwell, "Police Officers' Perceptions of Organizational Justice and Their Trust in the Public," *Police Practice and Research* 19 (4) (2018): 365–379; Maarten Van Craen and Wesley G. Skogan, "Achieving Fairness in Policing: The Link Between Internal and External Procedural Justice," *Police Quarterly* 20 (1) (2017): 3–23; and Tyler and Huo, *Trust in the Law*.

[78] Carr and Maxwell, "Police Officers' Perceptions of Organizational Justice and Their Trust in the Public"; Robert B. Cialdini, *Influence: The Psychology of Persuasion* (New York: Harper Business, 2021); Eberhardt, *Biased*; Gregory M. Walton and Alia J. Crum, *Handbook of Wise Interventions: How Social Psychology Can Help People Change* (New York: The Guilford Press, 2021); Van Craen and Skogan, "Achieving Fairness in Policing: The Link Between Internal and External Procedural Justice"; and Timothy D. Wilson, *Redirect: The Surprising New Science of Psychological Change* (Boston: Little, Brown and Company, 2011).

[79] Camp, Voigt, Hamedani, et al., "Leveraging Body-Worn Camera Footage to Assess the Effects of Training on Officer Communication During Traffic Stops."

[80] Winston and BondGraham, *The Riders Come Out at Night*.

[81] Hamedani, Markus, Hetey, and Eberhardt, "We Built This Culture (So We Can Change It): Seven Principles for Intentional Culture Change."

[82] Bryan Stevenson, *Just Mercy: A Story of Justice and Redemption* (New York: Spiegel & Grau, 2014).

[83] David DeBolt, "Federal Oversight of the Oakland Police Department May Be Nearing Its End, Attorneys Say," *The Oaklandside*, August 25, 2021, https://oaklandside.org/2021/08/25/federal-oversight-oakland-police-department-nearing-end-negotiated-settlement-agreement.

[84] Susie Neilson, "Oakland's Crime Rates Are Surging. Here's How They Compare with S.F. and Other Bay Area Cities," *San Francisco Chronicle*, August 4, 2023, https://www.sfchronicle.com/bayarea/article/oakland-bay-area-rates-18259788.php.

[85] Ibid.

[86] Oakland Police Department, *End of Year Crime Report – Citywide, 01 Jan. – 31 Dec., 2022* (Oakland: Oakland Police Department, 2023).

Disrupting the Effects of Implicit Bias: The Case of Discretion & Policing

Jack Glaser

Police departments tend to address operational challenges with training approaches, and implicit bias in policing is no exception. However, psychological scientists have found that implicit biases are very difficult to reduce in any lasting, meaningful way. Because they are difficult to change, and nearly impossible for the decision-maker to recognize, training to raise awareness or teach corrective strategies is unlikely to succeed. Recent empirical assessments of implicit bias trainings have shown, at best, no effect on racial disparities in officers' actions in the field. In the absence of effective training, a promising near-term approach for reducing racial disparities in policing is to reduce the frequency of actions most vulnerable to the influence of bias. Specifically, actions that allow relatively high discretion are most likely to be subject to bias-driven errors. Several cases across different policing domains reveal that when discretion is constrained in stop-and-search decisions, the impact of racial bias on searches markedly declines.

For anyone considering the topic of racial bias in policing, the murder of George Floyd, a Black man, by a White police officer in Minneapolis in 2020 looms large. The killing was slow (a nine-minute strangulation) and conducted in broad daylight. There were passionate, contemporaneous pleas from the victim and onlookers. One has to wonder if any amount of antibias training could have prevented that officer from killing Mr. Floyd. In contrast is the 2018 killing of Stephon Clark in his family's backyard in Sacramento. Clearly a wrongful killing by the police, the circumstances nevertheless differ considerably from the Floyd case. It was nighttime, and Clark, a twenty-two-year-old Black man, was shot to death by police officers who rushed around a blind corner, opening fire when they putatively mistook the phone in his hand for a gun.

As jarring as these accounts are, they are only two examples of a much larger problem revealed in the aggregate statistics. Prior to 2014 – the year a police officer fatally shot Michael Brown, an unarmed eighteen-year-old Black boy, in Ferguson, Missouri, and the widespread attention the subsequent protests garnered – data on fatal incidents of police use of force were sorely inadequate. While official statistics tended to put the count of fatal officer-involved shootings at roughly five hundred

https://doi.org/10.1162/daed_a_02053

per year in the United States, a thorough accounting by *The Washington Post* (corroborated by other organizations, like Fatal Encounters) has found that the actual number is roughly double that.[1] The racial disparities in these fatal events are marked. In a typical year, victims of these shootings are disproportionately Black, and the disparity is even greater among victims who were unarmed at the time of shooting.[2] Policy researcher Amanda Charbonneau and colleagues reported that, among off-duty police officers who were fatally shot by on-duty officers over a period studied, eight of ten were Black, a disproportion that we estimated had a less than one-in-a-million probability of occurring by chance.[3] Sociologists Frank Edwards, Hedwig Lee, and Michael Esposito used national statistics from 2013 to 2018 to estimate that the lifetime risk of being killed by police is about one in one thousand for Black men; twice the likelihood of American men overall.[4]

Fatal cases are just the tip of the iceberg. For nonfatal incidents, multiple research groups using heterogeneous methods have consistently found Black Americans to be disproportionately subject to all nonfatal levels of use of force by police.[5]

It is illuminating to further contrast the use of force and killings by police of unarmed Black men with what is, on its face, a more innocuous kind of police-civilian encounter, but one that happens with far greater frequency and has devastating cumulative effects on communities of color. These are discretionary investigative contacts, such as pedestrian and vehicle stops, many of which are based on vague pretexts like minor equipment violations or "furtive movements" that serve primarily to facilitate investigatory pat-downs or searches, most of which prove to be fruitless.[6] This essay considers the broad range of police-civilian encounters, from the routine to the deadly, because the implications for the role of implicit bias, and the promise of the available countermeasures, vary dramatically across the spectrum.

Implicit bias trainings are unlikely to make a difference for officers who will commit murder in cold blood. But for officers who are entering a fraught use-of-force situation (or, for that matter, are faced with the opportunity to prevent or de-escalate one), having a heightened awareness about the potential for bias-driven errors, and/or having an attenuated race-crime mental association, could make the difference in a consequential split-second decision. For officers engaged in more day-to-day policing, effective interventions might help them to focus their attention on operationally, ethically, and constitutionally valid indicators of criminal suspicion and opportunities to promote public safety.

Implicit bias is real, it is pervasive, and it matters. Implicit bias (also known as automatic bias or unconscious bias) refers to mental associations between social groups (such as races, genders) and characteristics (such as good/bad, aggressive) that are stored in memory outside of conscious awareness and are activated automatically and consequently skew judgments and affect behaviors of

individuals.[7] Other essays in this volume go into greater depth and breadth on the science and theory behind the concept of implicit bias, but I will provide here a succinct description that highlights the themes most important to efforts to disrupt implicit bias.[8]

The theoretical origins of implicit bias, a construct developed and widely used by social psychologists, are firmly planted in the sibling subfield of cognitive psychology. Cognitive psychologists interested in how people perceive, attend to, process, encode, store, and retrieve information used ingenious experimental methods to demonstrate that much of this information processing occurs outside of conscious awareness (implicitly) or control (automatically), enabling people to unknowingly, spontaneously, and effortlessly manage the voluminous flow of stimuli constantly passing through our senses.[9]

Beginning in the 1980s, social psychologists applied these theories and methods to understand how people process information about others, and in particular, with respect to the groups (racial, ethnic, gender, and so on) to which they belong.[10] This research area of implicit social cognition proved tremendously effective for demonstrating that people had mental associations about social categories (such as racial groups) that could be activated automatically, even if the holder of these associations consciously repudiated them. These associations could reflect stereotypes (associations between groups and traits or behavioral tendencies) or attitudes (associations between groups and negative or positive evaluation; that is, "prejudice").

A major advantage for the social science of intergroup bias provided by measures of implicit bias was that these methods could assess biases at a time when it was taboo to express them explicitly. At least as important, these methods measure biases people may not even know they hold and are unlikely to subjectively experience their activation or application, let alone effectively inhibit.

The methods for measuring implicit associations are indirect. In contrast to traditional methods for measuring beliefs and attitudes that involve asking people directly, or even subtler questionnaire approaches like the Modern Racism Scale, measures of implicit associations involve making inferences about the strength of the association.[11] That inference is based on the facility with which people process stimuli related to different categories, typically measured by the speed with which they respond to words or pictures that represent groups of people when they are paired with stimuli representing the category about which their association is being assessed.

In 1998, psychologists Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz published the first of very many reports of the Implicit Association Test (IAT).[12] The IAT is noteworthy because it is far and away the most widely utilized tool to assess implicit bias, and has benefited from the thorough exploration of its psychometric properties that has resulted. As described in detail by

Kate A. Ratliff and Colin Tucker Smith in this volume, the IAT yields a bias score that reflects the standardized average speed with which the participant responds when the categories are combined one way (for example, Black associated with good, White associated with bad) versus the other, thereby allowing for an inference that the individual associates one group with one trait (good or bad) more than the other.

Considering that the IAT is generating an index of the strength of someone's mental associations between categories based on the speed to press buttons in response to a disparate array of stimuli that are, by the way, presented in a different order for each participant, we do not expect it to be a *strong* predictor of anything; in scientific terms, it is "noisy," and should not be used for "diagnostic" purposes at the individual level. Nevertheless, when looking at aggregate data, the IAT and similar measures have been shown to have reasonably good construct validity and test-retest reliability.[13]

The IAT has become so influential, in part, because it has now been carried out literally millions of times through the Project Implicit website, which hosts numerous versions of the IAT that can be taken for demonstration or research purposes.[14] As a result, researchers have been able to test the convergent validity of the IAT, finding that it correlates reliably and predictably with explicit (that is, direct, questionnaire-based) measures of the same attitudes.[15] Therefore, although implicit bias scores are indirect, representing response speed differences to varied series of stimuli, they correlate with measures that, although subject to self-presentation bias, are clear on their face about what they are measuring.

More important than correspondence with explicit measures, which have their own limitations, implicit measures have been shown to correlate with behavior, specifically, discriminatory behavior.[16] Psychologist Benedek Kurdi and colleagues carried out a meta-analysis of over two hundred studies with tests of IAT-behavior relations, finding small but consistent positive relations, above and beyond (that is, after statistically controlling for) explicit measures of bias.[17] They also found that the more methodologically rigorous the study, the larger the relationship. Although these effects tend to be small, psychologists Anthony G. Greenwald, Mahzarin R. Banaji, and Brian A. Nosek, as well as legal scholar Jerry Kang in this issue, have rightly noted that small effects, when widespread and persistent, can have cumulatively large consequences.[18]

Some of these research findings involve correlations between implicit bias measures and highly important, real world discriminatory behaviors.[19] For example, economist Dan-Olof Rooth found that implicit preference for ethnically Swedish men over Arab-Muslim men in Sweden predicted the rate at which real firms invited applicants for interviews as a function of the ethnicity conveyed by the names on otherwise identical résumés: recruiters with stronger anti-Arab-Muslim implicit bias were less likely to invite applicants with Arab-Muslim

sounding names.[20] Implicit racial attitudes significantly predicted self-reported vote choice in the 2008 U.S. presidential election, even after controlling for common vote predictors like party identification, ideology, and race.[21] Implicit associations between the self and death/suicide predicted future suicide attempts in a psychiatric population.[22] In a sample of medical residents, implicit racial bias was associated with a decreased tendency to recommend an appropriate treatment for a Black patient, and an increased tendency for a White patient.[23]

In my own research, we have found that an implicit association between Black people and weapons (but not the generic Black-bad association) is a predictor of "shooter bias": the tendency to select a shoot (instead of a don't shoot) response when presented with an image of an armed Black man, as opposed to an armed White man.[24] Our study used a college undergraduate sample, but other studies of shooter bias have found it to be prevalent in police samples.[25]

Another line of research has found evidence of police officers taking longer to shoot Black individuals than White individuals, and being less likely to shoot unarmed Black people than unarmed White people.[26] In this simulation, officers were presented with video vignettes that lasted roughly forty seconds. In each vignette, the suspect appears early, but the decision to shoot, prompted by the appearance of a weapon, for example, occurs late. Under these conditions, officers may have time to marshal corrective strategies. In contrast, the shooter bias studies involve a series of rapid responses, with each trial taking less than one second. Interestingly, the same researchers who observed the reverse-racism effect in the more protracted simulation have found that police officers associate Black people and weapons, and that the association is most pronounced when they have had relatively little sleep.[27] Taken together, these sets of findings suggest that at least some officers would override their implicit racial bias if given the opportunity. This is consistent with the MODE (Motivation and Opportunity as Determinants) model of information processing used to explain attitude-behavior relationships.[28]

However, there is a well-established tradeoff between speed and accuracy when people make decisions.[29] Under realistic conditions, wherein there are distractions, distress, and a sense of threat, processing difficulties reflected in response latency are likely to translate into errors.

Be it in hiring, health care, voting, policing, or other consequential decision-making, implicit biases have been shown to be influential, implicating the need for effective interventions to promote nondiscrimination.

As cognitive psychologists demonstrated decades ago, implicit cognition is a constant fact of life. It serves an adaptive function of helping people manage a volume of information that would be impossible to handle consciously. It also helps us automatize the activation of memories and processes, such as driving a car, to free up conscious resources for more novel and com-

plex decisions.[30] This is true for memories of people and the categories in which we perceive them as belonging. As a consequence, implicit stereotypes and attitudes are pervasive. There is an extensive social psychological literature on what the sources and causes of these biases are, and there is a clear accounting of the extent of implicit bias from research using many thousands of IAT results gathered through Project Implicit.[31]

Directly relevant to the issue of implicit bias and policing, psychologists Eric Hehman, Jessica K. Flake, and Jimmy Calanchini have shown that regional variation in implicit racial bias (based on Project Implicit data) is associated with variation in racial disparities in police use of force, and psychologists Marleen Stelter, Iniobong Essien, Carsten Sander, and Juliane Degner have shown that county-level variation in both implicit and explicit prejudice is related to racial disparities in traffic stops.[32] The greater the average anti-Black prejudice, the greater the ratio of stops of Black people relative to their local population. These findings do not speak conclusively to whether there is a direct, causal link between police officers' implicit bias levels and their racially disparate treatment of community members. But they suggest that, at the very least, variation in the cultural milieu that gives rise to implicit biases affects police performance as well.

Given its prevalence and influence over important behaviors, there has long been interest in identifying conditions and methods for changing implicit biases. Cognitive social psychologists have been skeptical about prospects for meaningfully and lastingly changing implicit biases because of their very nature: they reflect well-learned associations that reside and are activated outside of our subjective experience and control. Furthermore, they would not serve their simplifying function well if they were highly subject to change. Being products of what we have encountered in our environments, implicit biases are unlikely to change without sustained shifts in the stimuli we regularly encounter. For that matter, even explicit attitudes and beliefs are difficult to modify.[33] Nevertheless, considerable exploration has been conducted of the conditions under which implicit biases can change, or at least fluctuate.

One important strain of research is on the malleability of implicit biases. Distinguishable from lasting change, malleability refers to contextual and strategic influences that can temporarily alter the manifestation of implicit biases, and considerable evidence has shown that the activation and application of implicit biases are far from inevitable. For example, social psychologist Nilanjana Dasgupta has found over a series of studies that scores on measures of implicit bias can be reduced (although rarely neutralized) by exposing people to positive examples or media representations from the disadvantaged group.[34] Social psychologist Irene V. Blair provided an early and compelling review of implicit bias malleability, noting that studies showed variation in implicit bias scores as a func-

tion of experimenter race and positive mental imagery, and weaker implicit stereotypes after extended stereotype negation training (that is, literally saying "no" to stereotype-consistent stimulus pairings).[35] On the other hand, there is research showing that implicit biases are highly resistant to change.[36] Recent efforts to examine the conditions under which implicit attitudes may or may not shift have revealed, for example, that evaluative statements are more impactful than repeated counter-attitudinal pairings, and that change is easier to achieve when associations are novel (in other words, learned in the lab) as opposed to preexisting.[37]

For my part, I have been interested in the possibility that egalitarian motivations can themselves operate implicitly, holding promise for automatic moderation of implicit bias effects.[38] Research has shown that goals and motives, like beliefs and attitudes, can operate outside of conscious awareness or control.[39] Furthermore, research on explicit prejudice has shown that motivation to control prejudiced responding, as measured with questionnaires, moderates the relation between implicit bias and expressed bias.[40] My colleagues and I developed a reaction time–based method to identify those who are most likely to be *implicitly* motivated to control prejudice (IMCP), finding that those who had a relatively strong implicit association between prejudice and badness (an implicit negative attitude toward prejudice) as well as a relatively strong association between themselves and prejudice (an implicit belief oneself is prejudiced) showed the weakest association between an implicit race-weapons stereotype and shooter bias.[41] We further found that only those high in our measure of IMCP were able to modulate their shooter bias when their cognitive resources were depleted, providing evidence that the motivation to control prejudice can be automatized (that is, operate largely independently of cognitive resources).[42]

Several robust efforts have been made to test for effective methods to lastingly reduce implicit bias. Social psychologists Patricia G. Devine, Patrick S. Forscher, Anthony J. Austin, and William T. L. Cox tested a multifaceted, long-duration program to "break the prejudice habit."[43] They developed an approach emphasizing the importance of people recognizing bias (awareness), being concerned about it (motivation), and having specific strategies for addressing it. Their program took place over an eight-week span as part of an undergraduate course, and they found significant reductions in (albeit, by no means elimination of) implicit bias four and eight weeks after the beginning of the program. However, a subsequent intervention experiment on gender bias among university faculty, while still showing promising effects on explicit and behavioral measures, did not replicate reductions in implicit bias.[44]

With respect to focused, short-term methods for reducing implicit bias, some extraordinarily systematic research has been conducted, finding that some approaches can partially reduce implicit racial bias, but that these effects are fleeting.[45] Social psychologist Calvin K. Lai and colleagues coordinated a "many labs"

collaboration to test a set of seventeen promising strategies to reduce implicit bias, specifically, the Black/White–bad/good association. The strategies include multiple methods to help participants engage with others' perspectives, expose them to counter-stereotypical examples, appeal to egalitarian values, recondition their evaluative associations, induce positive emotions, or provide ways to override biases. Additionally, an eighteenth strategy, "faking" the IAT, was tested. At least three research groups tested each strategy, allowing for statistically powerful, reliable inferences. While nine of these eighteen approaches yielded virtually no change in implicit bias as measured on the IAT, the other nine yielded statistically significant, albeit only partial, reductions. However, in a subsequent, careful, and robust study, Lai and colleagues retested the nine effective strategies, finding, first, that all were again able to cause statistically significant reductions in implicit bias, but that when the IAT was administered between two and twenty-four hours after the initial test, all but one of the groups' implicit bias scores had returned to baseline – the bias reduction effects were partial and short-lived.[46] Similarly, social psychologist Patrick S. Forscher and colleagues conducted a large meta-analysis of experiments testing methods to reduce scores on implicit bias measures, finding the typical effects to be weak.[47]

This is not by any means conclusive evidence that bias reduction strategies cannot have substantial, lasting effects, perhaps with the right dosing (duration and repetition). However, the body of evidence to date indicates that, without meaningful, lasting environmental change, implicit biases are resilient. This is entirely consistent with the theory and evidence regarding implicit cognition more generally: the ability to store, activate, and apply implicit memories automatically is adaptive. If implicit associations, particularly those well-learned (such as over a significant period of time), were highly malleable or changeable, they would not serve their function.

In policing, as in many other industries, providing trainings is a method of first resort when concerns about discrimination arise. Unfortunately, few of these trainings are accompanied by rigorous evaluations, let alone assessments including behavioral or performance outcomes.[48] Some systematic reviews of diversity trainings have found small effects on behavioral outcomes. Psychologist Zachary T. Kalinoski and colleagues found small- to medium-sized effects for "on-the-job behavior" in the six studies in their meta-analysis that included such behavioral outcomes.[49] In a large meta-analysis of diversity training program studies, psychologist Katerina Bezrukova and colleagues found relatively small effects on behavioral outcomes.[50] On the other hand, in their large-scale study, sociologists Alexandra Kalev, Frank Dobbin, and Erin Kelly found that diversity training had no effect on the racial or gender managerial composition of firms.[51] Psychologist Elizabeth Levy Paluck and colleagues have carefully reviewed the effects of diver-

sity trainings, finding few to have meaningful measures of behavioral outcomes, and for those few to be lacking evidence of effects on actual behavior.[52]

In policing, there has been considerable participation in diversity training, with much of it labeled as "implicit bias training," in particular. CBS News surveyed a sample of one hundred fifty-five large American municipal police departments, finding that 69 percent reported having carried out implicit bias trainings.[53] Departments and trainers, however, have not participated in robust evaluations of the effects of implicit bias training on officer performance, until recently. In 2018–2019, under the supervision of a court-appointed monitor resulting from a civil suit, one of the world's largest law enforcement agencies, the New York Police Department (NYPD), engaged the industry leader Fair and Impartial Policing in implicit bias training for its roughly thirty-six thousand sworn officers.[54] The effects of the training were evaluated effectively by exploiting the staggered rollout of the program, allowing for a comparison of field performance for officers before and after the training without confounding the comparison with any particular events that occurred simultaneously.[55] The researchers found that, while officers evaluated the training positively and reported greater understanding of the nature of implicit bias, only 27 percent reported attempting to apply their new training frequently (31 percent "sometimes") in the month following, while 42 percent reported not at all. More concerning, comparisons between pre- and post-training of the racial distributions of those stopped, frisked, searched, and who had force used against them revealed that, if anything, the percent who were Black increased. This study occurred from 2017 to 2019, a period after which the controversial stop-question-and-frisk (SQF) program had been ruled unconstitutional and dramatically reigned in, so disparities had already been somewhat reduced, leaving less room for improvement. However, as the study data reveal, while Black people – who make up about 25 percent of the city population – were 59.3 percent of those stopped in the first six months of 2019, they were only 47.6 percent of those arrested, suggesting that there remained considerable racial bias in who was being stopped.

A very recent, rigorous evaluation of the effects of another mainstream implicit bias training for police was conducted by Calvin K. Lai and Jaclyn A. Lisnek.[56] In this study, while trained officers indicated greater knowledge of bias that lasted at least one month after training, their increased concerns about bias, and understanding of the durability of bias, were more fleeting. With respect to behavioral outcomes, while officers indicated intentions to use strategies to manage bias following the training, their self-reported actual use of the strategies in the month after training was, disappointingly, lower than their self-reported use at baseline (prior to training).

The null effects on behavior come as no surprise to cognitive social psychologists, given that these trainings typically aim to, in a single day or less, mitigate the effects of cognitive biases that are learned over the lifespan, operate outside of

conscious awareness, and occur automatically. That said, implicit bias is not the only cause of discrimination, so it is especially discouraging that these trainings, which emphasize the importance of bias and awareness of it, do not appear to affect behavior through other channels, such as conscious, deliberate thought and behavior. On the other hand, this should not be all that surprising, given the very subtle and mixed effects of other forms of prejudice reduction trainings. This is not to say that implicit bias and other prejudice reduction trainings have no hope of meaningfully and lastingly reducing discrimination. There will need to be, however, further development and testing of training strategies that work. Until then, other avenues for disrupting implicit bias must also be explored.

In the absence of training that meaningfully and, ideally, lastingly reduces disparate treatment, a promising approach to reduce the impact of implicit biases is to constrain discretion. As Amanda Charbonneau and I, and others, have explained, police officers have a high degree of discretion (that is, latitude) in how they conduct their duties.[57] This stems in part from the vagueness of the regulatory standards, particularly "reasonable suspicion," that govern their practices. Suspicion is an inherently subjective experience, and its modifier "reasonable" is an intentionally vague standard that is often tautologically defined: "reasonable" is what a reasonable person (or officer) would think or do. Many people, in their professional endeavors, have discretion in how they carry out their jobs, including decisions about academic grading, admissions, and hiring; public- and private-sector hiring and promotions; legislative voting; public benefits eligibility; and mental and physical health care. Although individual professionals gain expertise through training and experience that may help them make good assessments, we rarely make decisions with complete information, and the evidence is clear that, in the absence of complete and specific information, we often rely on cognitive shortcuts like stereotypes, and/or interpret evidence in ways that are consistent with our prior conceptions or preferred outcomes.[58]

When discretion is high – for example, when decision-makers can use their own judgment in ambiguous situations – cognitive shortcuts like stereotypes have more opportunity to influence decisions. Analyses of real-world data on hiring and disciplinary decisions demonstrate that, in the absence of specific information, biases are influential. For example, economists Harry J. Holzer, Steven Raphael, and Michael A. Stoll found that employers who carried out criminal background checks were *more* likely to hire African Americans, suggesting that, in the absence of specific information about criminality, decision-makers may make the stereotype-consistent assumption.[59] With the specific information, they are less likely to discriminate. Similarly, economist Abigail Wozniak found that the implementation of legislation promoting drug testing resulted in substantial increases in Black employment rates, again raising the possibility that, in the absence of specific infor-

mation, stereotype-consistent judgments will disadvantage stigmatized groups in high-discretion decision-making like hiring, promotion, and retention.[60]

In the domain of school discipline, which bears important similarities and even a direct relationship to criminal justice (that is, the school-to-prison pipeline), psychologist Erik J. Girvan and colleagues Cody Gion, Kent McIntosh, and Keith Smolkowski found that, in a large dataset of school discipline cases, the vast majority of the variance in racial disparities was captured in high-discretion referrals.[61] Specifically, cases involving indicators of misconduct that were determined by the subjective assessment of school staff, as opposed to those with objective criteria, had far more racially disparate referral rates.

Specific to policing, Charbonneau and I have considered three large cases in which officer discretion can be operationalized in different ways.[62] We found that, across a range of law enforcement agencies, higher discretion in decisions to search was associated with greater disparities in search yield rates. Specifically, when discretion was high, White people who were searched were more likely to be found with contraband than were Black people or Latino people. In two of these cases (U.S. Customs and New York City), policy changes allow for a reasonably strong causal inference that reductions in discretion reduce disparities.

Comparisons of search yield rates (the percentage of searches that yield contraband) offer a compelling method to identify bias in law enforcement decisions. Drawing from the larger research literature on "outcome tests," the inference can be made that, if searches of one group of people are more likely to result in findings of contraband, then whatever is giving rise to decisions to search members of that group is generally a better indicator of criminal suspicion than whatever is triggering searches of other groups.[63] In other words, groups with higher search yield rates are probably being subjected to higher thresholds of suspicion in order to be searched. Groups who are searched based on lower levels of individual suspiciousness (perhaps because their group is stereotyped as prone to crime) will be less likely to be found in possession of evidence of crime. In turn, if one group has lower search yield rates than others, it can be inferred that there is group-based bias in at least some of the decisions to search. This could be compounded by group-based bias in decisions to surveil and stop, in the first place. When high discretion of who to surveil, stop, and search is afforded to officers, these decisions will be made under higher degrees of ambiguity (that is, less determined by codified criteria) and will therefore be more prone to the influence of biases such as racial stereotypes, thereby causing disparities.

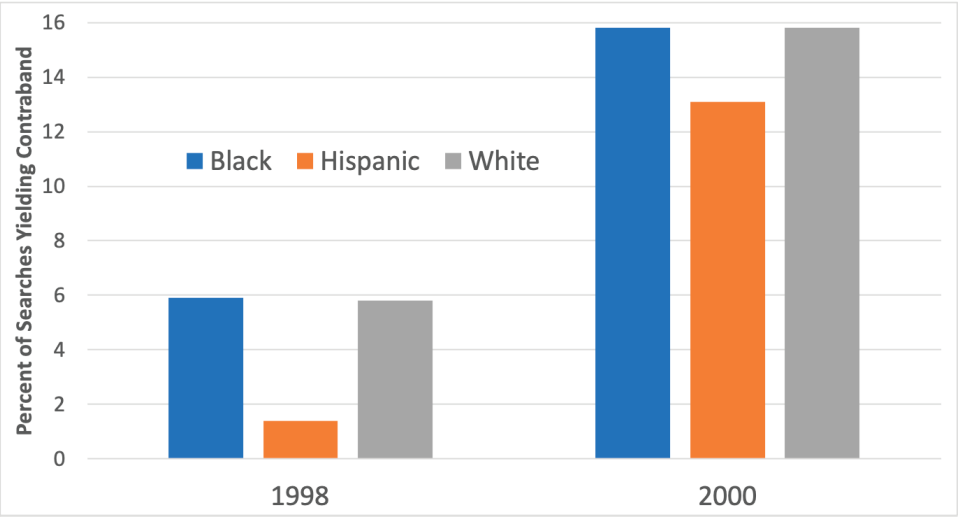
In 1999, the U.S. Customs Service (now Customs and Border Patrol) reduced the number of criteria for triggering a search of a traveler from forty-three to six, with the new criteria being more instrumentally related to smuggling.[64] Comparing the full year before to the full year after the reduction in search criteria, the number of searches declined 75 percent, but the search yield rate quadru-

pled. More important, the search yield rates became much less racially disparate. Prior to the change, the search yield rates for Hispanic people had been roughly one-quarter of the rate for Black people and White people who were searched, strongly suggesting that Hispanic people were being searched at lower thresholds of suspiciousness (because their searches were less likely to prove to be justified). As shown in Figure 1, after the reduction in search criteria, search yield rates increased overall, and nearly equalized across groups. Part of the inequity may have been due to a large share of Customs searches occurring at the U.S.-Mexico border, where searches may have been more frequent overall. However, the dramatic reduction in search yield disparities after the change in search criteria indicates that the disparity was mostly due to differential standards of suspicion being applied when discretion was high – when there were a lot of criteria to choose from. If the disparity had been due solely to different rates of searches at different ports of entry, the change in criteria would not have caused nearly as large a reduction in yield rate disparities.

The effects in the U.S. Customs case, in terms of increased yields overall and decreased disparities, are dramatic. This may be due in part to the nature of customs searches, which involve a decision (to search or not) about each person passing through the system, in contrast to searches in traffic enforcement or street policing, in which the decision to search is conditional upon the decisions to surveil and stop, which are also based on suspicion. In the noisier latter condition, the effects of discretion on search yield disparities would likely be smaller.

The second case involves the largest law enforcement agency in the United States. Over several decades, the NYPD has had an ebbing and flowing of the stop-question-and-frisk program, involving thousands of low-level stops of pedestrians with the primary goal of reducing street crime. A wave of increasing SQF began in the early 2000s, peaking in 2011 with over 685,000 stops in one year. About half of those stopped were Black people (mostly young men) – double their rate of residency – and about half of all of those stopped were subjected to frisks or searches, but the rate was considerably higher for Black people and Latino people than for White people. Officers most often recorded using highly subjective criteria, such as furtive movements, to justify their stops. Due to shifting political winds and a successful class action lawsuit, SQF declined (at least as indicated by reported stops) precipitously after 2011, plummeting to fewer than 20,000 stops per year by 2015.[65] As is commonly the case in search yield statistics, contraband and weapon discovery rates in 2011 were much higher for White people who were frisked than for Black people or Latino people. The White-Black yield rate ratio was 1.4-to-1 and 1.8-to-1 for contraband and weapons, respectively. In 2015, with a fraction of the number of stops, and the removal of furtive movements and other high-discretion reportable bases for stops, those ratios declined to 1.1-to-1 and 1.05-to-1, indicating that decisions to stop and frisk were less influenced by racial bias.

Figure 1
Percent of U.S. Customs Searches Yielding Contraband Before and After the 1999 Reduction in Search Criteria



Source: Author's image, based on data from Deborah Ramirez, Jennifer Hoopes, and Tara Lai Quinlan, "Defining Racial Profiling in a Post–September 11 World," *American Criminal Law Review* 40 (2003): 1195.

Especially telling are our analyses of statewide data from California, facilitated by the 2015 passage of the Racial and Identity Profiling Act (RIPA) requiring all law enforcement agencies in the state to report data on all traffic and pedestrian stops.[66] In contrast to the U.S. Customs and NYPD cases, where we compared racial disparities in search yield rates as a function of reduced discretion in search practices over time, with the RIPA data, we compared disparities across search types that varied in how discretionary they tend to be. For example, reviewing data from the first wave of RIPA – the eight largest departments in the state (including the Los Angeles Police Department, LA County Sheriff, and California Highway Patrol) – we found that yield rates were higher for White people than Black people and Latino people for searches based on supervision status (such as probation or parole), which allow officers considerable discretion, first to ask if someone is under supervision, and then to opt to search. However, for searches that were "procedural," such as those required during an arrest ("incident to arrest"), the search yield rates for White people were comparable to those for Black people and Latino people.[67]

Across these three cases, including a large, federal agency, an immense metropolitan police department, and the eight largest agencies in the most populous American state, we found that when officers' search discretion was relatively high, White people who were searched were more likely to be found in possession of contraband or weapons, indicating that White people were being subjected to higher thresholds of suspicion than Black people and Latino people in order to get stopped and/or searched. When discretion was relatively low (when search decisions were based on more stringent, prescribed criteria), yield rates were higher overall, and far less disparate. The evidence reviewed indicates that reducing discretion – in police stop-and-search practices, school discipline, private-sector hiring, and likely many other domains – is an effective method for reducing racial, ethnic, or other disparities. In the policing cases, at least, the overall improvements in search yield rates when discretion is low suggest that the effectiveness of the work need not be compromised. This was literally the case in Customs searches because, while searches dropped 75 percent, contraband discoveries quadrupled, resulting in roughly the same raw number of discoveries. That reductions in searches will have commensurate increases in yields is by no means likely, let alone guaranteed. This was certainly not the case in New York City, where the roughly 97 percent decline in pedestrian stops was accompanied by approximately a doubling in search yield rates. However, concerns that reducing SQF would result in an increase in crime were not borne out.[68] In fact, the continued decline in crime following SQF's near elimination was compelling enough to cause some rare public *mea culpas*.[69] It should also be noted that a large majority of the contraband recovered in NYPD searches was drug-related, while firearm seizures numbered in the hundreds, even at the peak of SQF. Even if high-discretion searches have, under some circumstances, a deterrent effect on crime, this must be weighed against the psychological harms caused by overpolicing, not to mention the violations of Fourth and Fourteenth Amendment protections against unreasonable searches and seizures and of equal protection.

When considering what can and cannot be done to disrupt the effects of implicit biases, it is crucial to bear in mind that implicit biases cause discriminatory judgments and actions *indirectly*. Because they operate outside of conscious awareness and control, and are generally not subjectively experienced by their holders, their effects are largely unintentional. Even an overt racist can have his bigotry enhanced (or possibly diminished) by implicit biases of which he is not aware.

An illustrative example of how implicit bias causes discrimination comes from a classic experiment that preceded the implicit bias innovations in psychological science. Psychologists John M. Darley and Paget H. Gross had research subjects evaluate the academic performance of a schoolgirl ostensibly named Hannah. Half of the sample was led to believe Hannah was from a low socioeconomic status

(SES) background, and the other half from a high SES background.[70] Splitting the sample yet again, half in each SES condition gave estimates of how they thought Hannah would do, while the other half rated her performance after watching a video of Hannah taking the tests. Among those who predicted Hannah's performance without watching the video, the low and high SES groups rated her about the same. Among those who actually observed her performance, even though all research participants watched the identical video, those who were given the impression that Hannah was low SES tended to rate her performance as below grade level, and those who were led to think she was high SES tended to rate her performance above grade level. They watched the same video, but interpreted the ambiguities in her performance in ways consistent with their stereotypes of low and high SES children. This was not intentional, or there would have been a similar pattern for those who did not see the video. People were, probably in good faith, doing their best to appraise Hannah's performance given the information they had. Their information about her socioeconomic status and the associated stereotypes skewed their perceptions. Likewise, implicit biases we may not even know we have, let alone endorse, can skew our perceptions and cause discriminatory judgments and behaviors.

This reality helps to explain how company hiring managers and staff will be inclined to interview people who have White- as opposed to Black-sounding names despite their résumés being identical, and why employers might tend to assume that Black applicants have criminal backgrounds or are drug users.[71] In the case of policing, officers are more likely to assume that people of color are involved in crime, even though searches of these individuals rarely bear this out, and they typically yield more evidence of criminality among White people who are searched – because the searches are biased.

In the absence of reliable methods for eliminating implicit (or, for that matter, explicit) biases, and with research indicating that trainings promoting cultural awareness, diversity, and fairness do not reliably reduce disparities in the real world, minimizing the vulnerability factors for discrimination is the best option. Reducing discretion and, ideally, replacing it with prescriptive guidance and systematic information (that is, valid criteria) has been shown to be effective with respect to stop-and-search decisions in policing.

Use of lethal force may require a special variant on the approach of reducing discretion. As discussed above with respect to demonstrations of police officers exhibiting "shooter bias" in simulations, situations in which police use force, and especially those that may involve lethal force, are fraught with vulnerabilities to errors. These situations typically involve time pressure, distraction, cognitive load, and intense emotions, including fear and anger. Many of these situations occur at night, adding visual ambiguity and heightening uncertainty and fear. As Jennifer T. Kubota describes, for many White Americans,

mere exposure to the image of a Black person's face triggers neurological activity consistent with fear, and the differential fear response to Black faces compared to White faces or neutral objects has been found to be associated with implicit racial bias.[72] This automatic fear response, occurring even in a mundane laboratory setting, is surely compounded by the anticipated (and often exaggerated) sense of mortal threat that police bring to civilian encounters.[73]

Given that implicit bias trainings for police, or even officers' self-reported utilization of trained strategies to interrupt bias, have been shown not to reduce disparate outcomes in stop, search, arrest, and use of nonlethal force, limiting the discretion with which police officers use force needs to be prioritized. In California, state law has been changed to require that lethal force be employed only when "necessary," a more stringent criterion than what it replaced: "reasonable."[74] However, it remains to be seen if this statutory change will translate into reduced levels of, and disparities in, excessive force, or if courts will merely apply a reasonableness standard to the necessity criterion (like what a "reasonable" officer would deem "necessary").

Some police departments appear to have had success in developing intensive trainings that reduce the unnecessary use of lethal force.[75] Practitioners emphasize the importance of officers slowing things down, keeping distance, and finding cover to reduce the likelihood of unnecessary force being used – approaches reflective of the challenges that the automatic activation of implicit racial bias presents. De-escalation training is also popular. But while there is at least one example of a police training program demonstrated to have reduced use of force and its collateral consequences, such as injuries, the evidence of trainings' effectiveness in general has been unclear.[76] To the extent that use of force is applied in a racially disparate manner, and the evidence of that is clear, reductions in unnecessary force should reduce disparities, just as reductions in unnecessary searches do.[77] Even if implicit bias is a substantial cause of disparities in police officers' use of force, interventions that directly target implicit bias are unlikely to succeed.

Researchers and practitioners can, and will, keep trying to look for practicable ways to reduce and/or override implicit biases through training. Some have made inroads, although the long-term effects on implicit biases themselves are tenuous, at best. While we wait for breakthroughs in methods and dosing, identifying institutional and personal vulnerabilities (such as hiring practices, enforcement practices, incentives, habits, distractions, cognitive load, and decision points) and possible methods to address them (for example, through constraints on discretion and prescriptions for better approaches) is more promising given the current state of the field. Rebecca C. Hetey, MarYam G. Hamedani, Hazel Rose Markus, and Jennifer L. Eberhardt describe prime examples of these kinds of prescriptive interventions in their contribution to this volume, including

requiring that officers provide more extensive explanations for their investigative stops.[78] As Hetey and coauthors as well as Manuel J. Galvan and B. Keith Payne argue in their essays in this issue, even if we could effectively disrupt implicit bias, we have to consider that structural factors such as historical inequities, incentives to punitiveness, and hierarchical institutional cultures are likely to be more influential than individual-level factors like implicit stereotyping. That said, individual and structural causes of discrimination are mutually reinforcing: structural inequities reinforce the negative attitudes, even at the implicit level, and vice versa.[79] Addressing structural factors can reduce considerable harm in the near future and, by attenuating disparities, possibly serve to soften individual-level biases, making them more conducive to change.

ABOUT THE AUTHOR

Jack Glaser is Professor at the Goldman School of Public Policy at the University of California, Berkeley. He is the author of *Suspect Race: Causes and Consequences of Racial Profiling* (2015), and has published in such journals as *Social and Personality Psychology Compass*, *Law & Human Behavior*, and *Policy Insights from Behavioral and Brain Sciences*.

ENDNOTES

1 "Police Shootings Database 2015–2023," *The Washington Post*, https://www.washingtonpost.com/graphics/investigations/police-shootings-database (accessed December 1, 2023).

2 As Charbonneau and colleagues report in 2015, the percentage of victims who were unarmed was more than twice as high for Black people than for White people, although that disparity declined in 2016, as the overall proportion of fatal officer-involved-shootings involving unarmed victims declined. Amanda Charbonneau, Katherine Spencer, and Jack Glaser, "Understanding Racial Disparities in Police Use of Lethal Force: Lessons from Fatal Police-on-Police Shootings," *Journal of Social Issues* 73 (4) (2017): 744–767.

3 Originally published in Christopher Stone, Zachary Carter, Thomas Belfiore, et al., *Reducing Inherent Danger: Report of the Task Force on Police-On-Police Shootings* (New York: New York State Task Force on Police-on-Police Shootings, 2010).

4 Frank Edwards, Hedwig Lee, and Michael Esposito, "Risk of Being Killed by Police Use of Force in the United States by Age, Race–Ethnicity, and Sex," *Proceedings of the National Academy of Sciences* 116 (34) (2019): 16793–16798.

5 For example, Roland G. Fryer Jr., "An Empirical Analysis of Racial Differences in Police Use of Force," *Journal of Political Economy* 127 (3) (2019): 1210–1261; Amanda Geller, Phillip Atiba Goff, Tracey Lloyd, et al., "Measuring Racial Disparities in Police Use of Force: Methods Matter," *Journal of Quantitative Criminology* 37 (2021): 1083–1113.

6 Jack Glaser, *Suspect Race: Causes and Consequences of Racial Profiling* (Oxford: Oxford University Press, 2015).

[7] Anthony G. Greenwald and Mahzarin R. Banaji, "Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes," *Psychological Review* 102 (1) (1995): 4.

[8] Kirsten N. Morehouse and Mahzarin R. Banaji, "The Science of Implicit Race Bias: Evidence from the Implicit Association Test," *Dædalus* 153 (1) (Winter 2024): 21–50, https://www.amacad.org/publication/science-implicit-race-bias-evidence-implicit-association-test; Kate A. Ratliff and Colin Tucker Smith, "The Implicit Association Test," *Dædalus* 153 (1) (Winter 2024): 51–64, https://www.amacad.org/publication/implicit-association-test; and Rebecca C. Hetey, MarYam G. Hamedani, Hazel Rose Markus, and Jennifer L. Eberhardt, "'When the Cruiser Lights Come On': Using the Science of Bias & Culture to Combat Racial Disparities in Policing," *Dædalus* 153 (1) (Winter 2024): 123–150, https://www.amacad.org/publication/when-cruiser-lights-come-using-science-bias-culture-combat-racial-disparities-policing.

[9] For example, David E. Meyer and Roger W. Schvaneveldt, "Meaning, Memory Structure, and Mental Processes: People's Rapid Reactions to Words Help Reveal How Stored Semantic Information Is Retrieved," *Science* 192 (4234) (1976): 27–33; and James H. Neely, "Semantic Priming and Retrieval from Lexical Memory: Roles of Inhibitionless Spreading Activation and Limited-Capacity Attention," *Journal of Experimental Psychology: General* 106 (3) (1977): 226.

[10] Patricia G. Devine, "Stereotypes and Prejudice: Their Automatic and Controlled Components," *Journal of Personality and Social Psychology* 56 (1) (1989): 5; and John F. Dovidio, Nancy Evans, and Richard B. Tyler, "Racial Stereotypes: The Contents of Their Cognitive Representations," *Journal of Experimental Social Psychology* 22 (1) (1986): 22–37.

[11] John B. McConahay, Betty B. Hardee, and Valerie Batts, "Modern Racism Scale," *Personality and Social Psychology Bulletin* (1980).

[12] Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz, "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test," *Journal of Personality and Social Psychology* 74 (6) (1998): 1464–1480.

[13] William A. Cunningham, Kristopher J. Preacher, and Mahzarin R. Banaji, "Implicit Attitude Measures: Consistency, Stability, and Convergent Validity," *Psychological Science* 12 (2) (2001): 163–170.

[14] In this volume, see Morehouse and Banaji, "The Science of Implicit Race Bias."

[15] Brian A. Nosek, "Moderators of the Relationship between Implicit and Explicit Evaluation," *Journal of Experimental Psychology: General* 134 (4) (2005): 565; Wilhelm Hofmann, Bertram Gawronski, Tobias Gschwendtner, et al., "A Meta-Analysis on the Correlation between the Implicit Association Test and Explicit Self-Report Measures," *Personality and Social Psychology Bulletin* 31 (10) (2005): 1369–1385; and Anthony G. Greenwald, T. Andrew Poehlman, Eric Luis Uhlmann, and Mahzarin R. Banaji, "Understanding and Using the Implicit Association Test: III. Meta-Analysis of Predictive Validity," *Journal of Personality and Social Psychology* 97 (1) (2009): 17.

[16] For parallel with the term "explicit measures," and reflecting common usage, I will sometimes use the term "implicit measure," but recognize that these methods are more accurately described as *indirect measures of implicit bias* because in many cases, it is evident to participants what the task is measuring. The bias being measured is implicit, in the sense that it is not consciously accessible, but the measure can be downright obtrusive.

[17] Benedek Kurdi, Allison E. Seitchik, Jordan R. Axt, et al., "Relationship between the Implicit Association Test and Intergroup Behavior: A Meta-Analysis," *American Psychologist* 74 (5) (2019): 569–586.

[18] Anthony G. Greenwald, Mahzarin R. Banaji, and Brian A. Nosek, "Statistically Small Effects of the Implicit Association Test Can Have Societally Large Effects," *Journal of Personality and Social Psychology* 108 (4) (2015): 553–561; and Jerry Kang, "Little Things Matter a Lot: The Significance of Implicit Bias, Practically & Legally," *Dædalus* 153 (1) (Winter 2024): 193–212, https://www.amacad.org/publication/little-things-matter-lot-significance-implicit-bias-practically-legally.

[19] For a review, see John T. Jost, Laurie A. Rudman, Irene V. Blair, et al., "The Existence of Implicit Bias Is Beyond Reasonable Doubt: A Refutation of Ideological and Methodological Objections and Executive Summary of Ten Studies That No Manager Should Ignore," *Research in Organizational Behavior* 29 (2009): 39–69.

[20] Dan-Olof Rooth, "Automatic Associations and Discrimination in Hiring: Real World Evidence," *Labour Economics* 17 (3) (2010): 523–534.

[21] Christopher Finn and Jack Glaser, "Voter Affect and the 2008 U.S. Presidential Election: Hope and Race Mattered," *Analyses of Social Issues and Public Policy* 10 (1) (2010): 262–275.

[22] Matthew K. Nock, Jennifer M. Park, Christine T. Finn, et al., "Measuring the Suicidal Mind: Implicit Cognition Predicts Suicidal Behavior," *Psychological Science* 21 (4) (2010): 511–517.

[23] Alexander R. Green, Dana R. Carney, Daniel J. Pallin, et al., "Implicit Bias among Physicians and Its Prediction of Thrombolysis Decisions for Black and White Patients," *Journal of General Internal Medicine* 22 (9) (2007): 1231–1238.

[24] Jack Glaser and Eric D. Knowles, "Implicit Motivation to Control Prejudice," *Journal of Experimental Social Psychology* 44 (1) (2008): 164–172; and Joshua Correll, Bernadette Park, Charles M. Judd, and Bernd Wittenbrink, "The Police Officer's Dilemma: Using Ethnicity to Disambiguate Potentially Threatening Individuals," *Journal of Personality and Social Psychology* 83 (6) (2002): 1314–1329.

[25] Joshua Correll, Bernadette Park, Charles M. Judd, et al., "Across the Thin Blue Line: Police Officers and Racial Bias in the Decision to Shoot," *Journal of Personality and Social Psychology* 92 (6) (2007): 1006–1023; and E. Ashby Plant and B. Michelle Peruche, "The Consequences of Race for Police Officers' Responses to Criminal Suspects," *Psychological Science* 16 (3) (2005): 180–183.

[26] Lois James, Stephen M. James, and Bryan J. Vila, "The Reverse Racism Effect: Are Cops More Hesitant to Shoot Black than White Suspects?" *Criminology & Public Policy* 15 (2) (2016): 457–479.

[27] Lois James, "The Stability of Implicit Racial Bias in Police Officers," *Police Quarterly* 21 (1) (2018): 30–52.

[28] Michael A. Olson and Russell H. Fazio, "Implicit and Explicit Measures of Attitudes: The Perspective of the MODE Model," in *Attitudes: Insights from the New Implicit Measures*, ed. Richard E. Petty, Russel H. Fazio, and Pablo Briñol (New York: Psychology Press, 2008), 39–84.

[29] Robert Sessions Woodworth, "Accuracy of Voluntary Movement," *The Psychological Review: Monograph Supplements* 3 (3) (1899): i–144.

[30] John A. Bargh and Tanya L. Chartrand, "The Unbearable Automaticity of Being," *American Psychologist* 54 (7) (1999): 462–479; and Susan T. Fiske and Shelley E. Taylor, *Social Cognition* (Boston: McGraw-Hill Book Company, 1991).

[31] Laurie A. Rudman, "Sources of Implicit Attitudes," *Current Directions in Psychological Science* 13 (2) (2004): 79–82; Tessa E. S. Charlesworth and Mahzarin R. Banaji, "Patterns of Implicit and Explicit Attitudes: I. Long-Term Change and Stability from 2007 to 2016," *Psychological Science* 30 (2) (2019): 174–192; and Brian A. Nosek, Frederick L. Smyth, Jeffrey J. Hansen, et al., "Pervasiveness and Correlates of Implicit Attitudes and Stereotypes," *European Review of Social Psychology* 18 (1) (2007): 36–88. See also, in this volume, Morehouse and Banaji, "The Science of Implicit Race Bias."

[32] Eric Hehman, Jessica K. Flake, and Jimmy Calanchini, "Disproportionate Use of Lethal Force in Policing Is Associated with Regional Racial Biases of Residents," *Social Psychological and Personality Science* 9 (4) (2018): 393–401; and Marleen Stelter, Iniobong Essien, Carsten Sander, and Juliane Degner, "Racial Bias in Police Traffic Stops: White Residents' County-Level Prejudice and Stereotypes Are Related to Disproportionate Stopping of Black Drivers," *Psychological Science* 33 (4) (2022): 483–496.

[33] Alice H. Eagly and Shelly Chaiken, "Attitude Strength, Attitude Structure, and Resistance to Change," *Attitude Strength: Antecedents and Consequences* 4 (2) (1995): 413–432.

[34] Nilanjana Dasgupta, "Implicit Attitudes and Beliefs Adapt to Situations: A Decade of Research on the Malleability of Implicit Prejudice, Stereotypes, and the Self-Concept," *Advances in Experimental Social Psychology* 47 (2013): 233–279.

[35] Irene V. Blair, "The Malleability of Automatic Stereotypes and Prejudice," *Personality and Social Psychology Review* 6 (3) (2002): 242–261; Brian S. Lowery, Curtis D. Hardin, and Stacey Sinclair, "Social Influence Effects on Automatic Racial Prejudice," *Journal of Personality and Social Psychology* 81 (5) (2001): 842; Irene V. Blair, Jennifer E. Ma, and Alison P. Lenton, "Imagining Stereotypes Away: The Moderation of Implicit Stereotypes through Mental Imagery," *Journal of Personality and Social Psychology* 81 (5) (2001): 828–841; and Kerry Kawakami, John F. Dovidio, Jasper Moll, Sander Hermsen, and Abby Russin, "Just Say No (To Stereotyping): Effects of Training in the Negation of Stereotypic Associations on Stereotype Activation," *Journal of Personality and Social Psychology* 78 (5) (2000): 871–888.

[36] Aiden P. Gregg, Beate Seibt, and Mahzarin R. Banaji, "Easier Done than Undone: Asymmetry in the Malleability of Implicit Preferences," *Journal of Personality and Social Psychology* 90 (1) (2006): 1–20; and Jennifer A. Joy-Gaba and Brian A. Nosek, "The Surprisingly Limited Malleability of Implicit Racial Evaluations," *Social Psychology* 41 (3) (2010): 137–146.

[37] Benedek Kurdi and Mahzarin R. Banaji, "Repeated Evaluative Pairings and Evaluative Statements: How Effectively Do They Shift Implicit Attitudes?" *Journal of Experimental Psychology: General* 146 (2) (2017): 194–213; and Benedek Kurdi, Kirsten N. Morehouse, and Yarrow Dunham, "How Do Explicit and Implicit Evaluations Shift? A Preregistered Meta-Analysis of the Effects of Co-Occurrence and Relational Information," *Journal of Personality and Social Psychology* 124 (6) (2022): 1174–1202.

[38] Jack Glaser and Mahzarin R. Banaji, "When Fair Is Foul and Foul Is Fair: Reverse Priming in Automatic Evaluation," *Journal of Personality and Social Psychology* 77 (4) (1999): 669–687; Jack Glaser and John F. Kihlstrom, "Compensatory Automaticity: Unconscious Volition Is Not an Oxymoron," *The New Unconscious* (2005): 171–195; and Gordon B. Moskowitz, Peter M. Gollwitzer, Wolfgang Wasel, and Bernd Schaal, "Preconscious Control of Stereotype Activation through Chronic Egalitarian Goals," *Journal of Personality and Social Psychology* 77 (1) (1999): 167–184.

[39] Tanya L. Chartrand and John A. Bargh, "Automatic Activation of Impression Formation and Memorization Goals: Nonconscious Goal Priming Reproduces Effects of Explicit Task Instructions," *Journal of Personality and Social Psychology* 71 (3) (1996): 464; and James Y. Shah and Arie W. Kruglanski, "When Opportunity Knocks: Bottom-Up Priming of Goals by Means and Its Effects on Self-Regulation," *Journal of Personality and Social Psychology* 84 (6) (2003): 1109.

[40] Russell H. Fazio, Joni R. Jackson, Bridget C. Dunton, and Carol J. Williams, "Variability in Automatic Activation as an Unobtrusive Measure of Racial Attitudes: A Bona Fide Pipeline?" *Journal of Personality and Social Psychology* 69 (6) (1995): 1013; and E. Ashby Plant and Patricia G. Devine, "Internal and External Motivation to Respond without Prejudice," *Journal of Personality and Social Psychology* 75 (3) (1998): 811.

[41] Glaser and Knowles, "Implicit Motivation to Control Prejudice."

[42] Sang Hee Park, Jack Glaser, and Eric D. Knowles, "Implicit Motivation to Control Prejudice Moderates the Effect of Cognitive Depletion on Unintended Discrimination," *Social Cognition* 26 (4) (2008): 401–419.

[43] Patricia G. Devine, Patrick S. Forscher, Anthony J. Austin, and William T. L. Cox, "Long-Term Reduction in Implicit Race Bias: A Prejudice Habit-Breaking Intervention," *Journal of Experimental Social Psychology* 48 (6) (2012): 1267–1278.

[44] Patrick S. Forscher, Chelsea Mitamura, Emily L. Dix, et al., "Breaking the Prejudice Habit: Mechanisms, Timecourse, and Longevity," *Journal of Experimental Social Psychology* 72 (2017): 133–146.

[45] Calvin K. Lai, Maddalena Marini, Steven A. Lehr, et al., "Reducing Implicit Racial Preferences: I. A Comparative Investigation of 17 Interventions," *Journal of Experimental Psychology: General* 143 (4) (2014): 1765.

[46] One of the strategies, "using implementation intentions," showed a barely significant reduced bias at the second assessment; and this effect was much reduced from the first assessment. Calvin K. Lai, Allison L. Skinner, Erin Cooley, et al., "Reducing Implicit Racial Preferences: II. Intervention Effectiveness across Time," *Journal of Experimental Psychology: General* 145 (8) (2016): 1001.

[47] Patrick S. Forscher, Calvin K. Lai, Jordan R. Axt, et al., "A Meta-Analysis of Procedures to Change Implicit Measures," *Journal of Personality and Social Psychology* 117 (3) (2019): 522.

[48] Elizabeth Levy Paluck and Donald P. Green, "Prejudice Reduction: What Works? A Review and Assessment of Research and Practice," *Annual Review of Psychology* 60 (2009): 339–367.

[49] Zachary T. Kalinoski, Debra Steele-Johnson, Elizabeth J. Peyton, et al., "A Meta-Analytic Evaluation of Diversity Training Outcomes," *Journal of Organizational Behavior* 34 (8) (2013): 1076–1104.

[50] Katerina Bezrukova, Chester S. Spell, Jamie L. Perry, and Karen A. Jehn, "A Meta-Analytical Integration of Over 40 Years of Research on Diversity Training Evaluation," *Psychological Bulletin* 142 (11) (2016): 1227.

[51] Alexandra Kalev, Frank Dobbin, and Erin Kelly, "Best Practices or Best Guesses? Assessing the Efficacy of Corporate Affirmative Action and Diversity Policies," *American Sociological Review* 71 (4) (2006): 589–617. See also Alexandra Kalev and Frank Dobbin, "Retooling Career Systems to Fight Workplace Bias: Evidence from U.S. Corporations,"

Dædalus 153 (1) (Winter 2024): 213–230, https://www.amacad.org/publication/retooling-career-systems-fight-workplace-bias-evidence-us-corporations.

[52] Elizabeth Levy Paluck, Roni Porat, Chelsey S. Clark, and Donald P. Green, "Prejudice Reduction: Progress and Challenges," *Annual Review of Psychology* 72 (2021): 533–560.

[53] CBS News, "We Asked 155 Police Departments about Their Racial Bias Training. Here's What They Told Us," August 7, 2019, https://www.cbsnews.com/news/racial-bias-training-de-escalation-training-policing-in-america.

[54] Disclosure: I have served as a paid expert for the lead plaintiffs in the New York City lawsuit.

[55] Robert E. Worden, Sarah J. McLean, Robin S. Engel, et al., *The Impacts of Implicit Bias Awareness Training in the NYPD* (Cincinnati: The John F. Finn Institute for Public Safety, Inc. and The Center for Police Research and Policy at the University of Cincinnati, 2020).

[56] Calvin K. Lai and Jaclyn A. Lisnek, "The Impact of Implicit-Bias-Oriented Diversity Training on Police Officers' Beliefs, Motivations, and Actions," *Psychological Science* 34 (4) (2023): 424–434.

[57] Amanda Charbonneau and Jack Glaser, "Suspicion and Discretion in Policing: How Laws and Policies Contribute to Inequity," *UC Irvine Law Review* 11 (2020): 1327; and Herman Goldstein, "Police Discretion: The Ideal Versus the Real," *Public Administration Review* 23 (3) (1963): 140–148.

[58] Charles G. Lord, Lee Ross, and Mark R. Lepper, "Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence," *Journal of Personality and Social Psychology* 37 (11) (1979): 2098.

[59] Harry J. Holzer, Steven Raphael, and Michael A. Stoll, "Perceived Criminality, Criminal Background Checks, and the Racial Hiring Practices of Employers," *The Journal of Law and Economics* 49 (2) (2006): 451–480.

[60] Abigail Wozniak, "Discrimination and the Effects of Drug Testing on Black Employment," *Review of Economics and Statistics* 97 (3) (2015): 548–566.

[61] Russell J. Skiba, Mariella I. Arredondo, and Natasha T. Williams, "More than a Metaphor: The Contribution of Exclusionary Discipline to a School-to-Prison Pipeline," *Equity & Excellence in Education* 47 (4) (2014): 546–564; and Erik J. Girvan, Cody Gion, Kent McIntosh, and Keith Smolkowski, "The Relative Contribution of Subjective Office Referrals to Racial Disproportionality in School Discipline," *School Psychology Quarterly* 32 (3) (2017): 392.

[62] Charbonneau and Glaser, "Suspicion and Discretion in Policing."

[63] Ian Ayres, "Outcome Tests of Racial Disparities in Police Practices," *Justice Research and Policy* 4 (1–2) (2002): 131–142.

[64] Deborah Ramirez, Jennifer Hoopes, and Tara Lai Quinlan, "Defining Racial Profiling in a Post–September 11 World," *American Criminal Law Review* 40 (2003): 1195; and Malcolm Gladwell, "Troublemakers: What Pitbulls Can Teach Us about Profiling," *The New Yorker*, January 29, 2006.

[65] Disclosure: I have served as a paid expert for the Floyd plaintiffs during the remedy phase of the NYPD SQF lawsuit.

[66] Disclosure: I have served as a paid expert consultant to the California Department of Justice, advising on the analysis and interpretation of RIPA data.

[67] Charbonneau and Glaser, "Suspicion and Discretion in Policing."

[68] John MacDonald, Jeffrey Fagan, and Amanda Geller, "The Effects of Local Police Surges on Crime and Arrests in New York City," *PLOS ONE* 11 (6) (2016): e0157223.

[69] Kyle Smith, "We Were Wrong about Stop and Frisk," *The National Review*, January 1, 2018, https://www.nationalreview.com/2018/01/new-york-city-stop-and-frisk-crime-decline-conservatives-wrong; and Anthony Fisher, "*NY Daily News* Admits 'We Were Wrong' about Stop and Frisk," *Reason Magazine*, August 9, 2016.

[70] John M. Darley and Paget H. Gross, "A Hypothesis-Confirming Bias in Labeling Effects," *Journal of Personality and Social Psychology* 44 (1) (1983): 20.

[71] Marianne Bertrand and Sendhil Mullainathan, "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review* 94 (4) (2004): 991–1013; Holzer, Raphael, and Stoll, "Perceived Criminality, Criminal Background Checks, and the Racial Hiring Practices of Employers"; and Wozniak, "Discrimination and the Effects of Drug Testing on Black Employment."

[72] Jennifer T. Kubota, "Uncovering Implicit Racial Bias in the Brain: The Past, Present & Future," *Dædalus* 153 (1) (Winter 2024): 84–105, https://www.amacad.org/publication/uncovering-implicit-racial-bias-brain-past-present-future; and Elizabeth A. Phelps, Kevin J. O'Connor, William A. Cunningham, et al., "Performance on Indirect Measures of Race Evaluation Predicts Amygdala Activation," *Journal of Cognitive Neuroscience* 12 (5) (2000): 729–738.

[73] Michael Sierra-Arévalo, "American Policing and the Danger Imperative," *Law & Society Review* 55 (1) (2021): 70–103.

[74] Worden, McLean, Engel, et al., "The Impacts of Implicit Bias Awareness Training in the NYPD"; Lai and Lisnek, "The Impact of Implicit-Bias-Oriented Diversity Training on Police Officers' Beliefs, Motivations, and Actions"; and California Assembly Bill 392 (2019).

[75] Robert Rogers, "Use of Deadly Force by Police Disappears on Richmond Streets," *East Bay Times*, September 6, 2014.

[76] Robin S. Engel, Nicholas Corsaro, Gabrielle T. Isaza, and Hannah D. McManus, "Assessing the Impact of De-Escalation Training on Police Behavior: Reducing Police Use of Force in the Louisville, KY Metro Police Department," *Criminology & Public Policy* 21 (2) (2022): 199–233; and Robin S. Engel, Hannah D. McManus, and Tamara D. Herold, "Does De-Escalation Training Work? A Systematic Review and Call for Evidence in Police Use-of-Force Reform," *Criminology & Public Policy* 19 (3) (2020): 721–759.

[77] Geller, Goff, Lloyd, et al., "Measuring Racial Disparities in Police Use of Force."

[78] Hetey, Hamedani, Markus, and Eberhardt, "'When the Cruiser Lights Come On.'"

[79] Ibid.; Manuel J. Galvan and B. Keith Payne, "Implicit Bias as a Cognitive Manifestation of Systemic Racism," *Dædalus* 153 (1) (Winter 2024): 106–122, https://www.amacad.org/publication/disrupting-effects-implicit-bias-case-discretion-policing; and John T. Jost and Mahzarin R. Banaji, "The Role of Stereotyping in System-Justification and the Production of False Consciousness," *British Journal of Social Psychology* 33 (1) (1994): 1–27.

Roles for Implicit Bias Science in Antidiscrimination Law

Anthony G. Greenwald & Thomas Newkirk

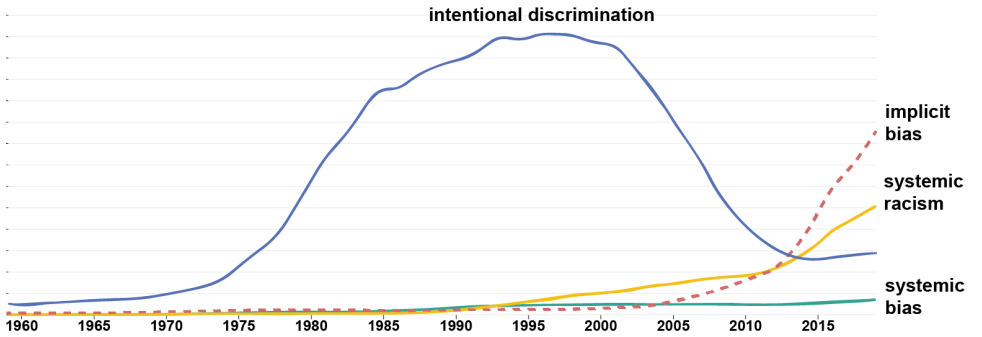
Declining scholarly interest in intentional discrimination may be due to rapid growth of interest in systemic biases and implicit biases. Systemic biases are produced by organizational personnel doing their assigned jobs, but nevertheless causing adverse impacts to members of protected classes as identified in civil rights laws. Implicit biases are culturally formed stereotypes and attitudes that cause selective harms to protected classes while operating mostly outside of conscious awareness. Both are far more pervasive and responsible for much greater adversity than caused by overt, explicit bias, such as hate speech. Scientific developments may eventually influence jurisprudence to reduce effects of systemic and implicit biases, but likely not rapidly. We conclude by describing possibilities for executive leadership in both public and private sectors to ameliorate discrimination faster and more effectively than is presently likely via courts and legislation.

Scholarly and scientific understanding of discrimination have developed greatly since implicit bias was introduced almost thirty years ago. Figure 1 illustrates the usage frequency of four discrimination-related terms that appeared in English-language books from 1959 to 2019. The plot reveals a long dominance of *intentional discrimination*, peaking in the first decade of the twenty-first century, followed by a more recent decline. Two terms rose to prominence in only the last twenty years, surpassing intentional discrimination by 2013: *implicit bias* and *systemic racism*. These trends signal a rapid societal assimilation of recent work by social scientists, psychological scientists, and legal scholars.

Implicit biases are a subset of one's social knowledge. They include mental associations that are the core of attitudes and stereotypes, acquired continuously, starting early in life. These associations are triggered automatically and without one's awareness during encounters with members of the demographic groups with which they are associated. When activated, the associated attitudes and stereotypes influence thoughts, judgment, and behavior that may thereby be biased toward or against members of those demographic groups. Implicit bias contrasts with explicit bias, a widely used label for consciously accessible beliefs that serve as a basis for (quite possibly) biased judgments and decisions.[1]

https://doi.org/10.1162/daed_a_02054

Figure 1
Four Discrimination-Related Terms in English-Language Books, 1959–2019



Usage, from 1959 through 2019, of four concepts prominent in English-language scholarly treatments of intergroup discrimination. Source: This plot was produced in Google Ngram by entering the four two-word terms, case insensitive, separated by commas, into the Ngram Viewer's search box, with smoothing set at three years. Google Ngram Viewer, https://books.google.com/ngrams/graph?content=intentional+discrimination%2Cimplicit+bias%2Csystemic+racism%2Csystemic+bias&year_start=1960&year_end=2015&corpus=en-2019&smoothing=3# (accessed December 15, 2023).

Systemic bias is a term we use in place of *systemic racism*, even though the latter term has had much more active use by legal scholars and social scientists since the 1980s (see Figure 1). We avoid using systemic racism both because systemic bias is not limited to race and because the *-ism* suffix connotes a negative mental attitude that is not a component of most of the phenomena now taken to exemplify systemic racism. Systemic biases are rooted in bureaucratic practices that are not in the human mind, but are codified in, among other places, corporate manuals and legislated regulations.

Both implicit biases and systemic biases can produce discrimination that occurs as intentional behavior, and both can occur, when not accompanied by explicit bias, without intent to harm. We are among a growing proportion of scholars and scientists who understand that, in combination (and likely also separately), implicit and systemic biases account for substantially more discriminatory harm than is due to explicit biases.

There are four empirically established properties of implicit biases, each with its own particular challenges: pervasiveness, predictive validity, lack of awareness, and resistance to change.

Pervasiveness. Multiple large studies using the Implicit Association Test (IAT)[2] have found that implicit biases are evident in many people. In this volume, Kirsten N. Morehouse and Mahzarin R. Banaji present in detail the evidence for the pervasiveness of implicit race bias, as measured by the IAT metric of racial preference for White relative to Black, a measure often identified as revealing "automatic White preference."[3] Combining data over fourteen years (2007 – 2020), Morehouse and Banaji observe that "2.1 of 3.3 million respondents automatically associated the attribute 'Good' (relative to 'Bad') more so with White than Black Americans."[4] By contrast, on a self-report measure of explicit White preference, only 29 percent preferred White relative to Black and "60 percent of respondents reported equal liking for both groups."[5] Data from many volunteers' performance on IAT measures have accumulated at the Project Implicit website, where visitors can choose to complete any of more than a dozen IAT measures of intergroup attitudes or stereotypes.[6] Visitors' performances on these IATs typically reveal that implicit biases are both stronger and more widely prevalent than explicit biases for measures concerning old compared with young, abled compared with disabled, gay compared with straight, male compared with female, Native American compared with White American, light skinned compared with dark skinned, thin compared with fat, and European American compared with Asian American. Numerous other attitude and trait dimensions have been tested and described in research publications, similarly often showing greater prevalence of implicit than explicit biases, but without numbers of respondents approaching the very large proportion of completed tests obtained and archived at the Project Implicit website.

Predictive validity. Discriminatory behavior is reliably predicted by IAT measures of implicit biases. Three meta-analyses of predictive validity of IAT measures have supported this conclusion.[7] It is not presently possible (nor will it likely be in the foreseeable future) to conduct true experimental tests that could establish the interpretation that implicit bias is a *cause* of discriminatory behavior. On the other hand, as an explanation for the observed correlations of implicit bias measures with discriminatory behavior, this causal interpretation has only one competitor, which is that implicit biases and discriminatory behavior have shared causes. At present, and also for the foreseeable future, there is no practical method of using either laboratory or field experimentation to choose between the implicit-bias-as-cause theory and the shared-causes theory.[8] It is therefore reasonable to treat implicit bias either as itself a cause of discrimination or as an indicator of a not-yet-identified precursor of both IAT-measured bias and the discriminatory behavior measures with which IAT measures are found to be correlated.

Lack of awareness. Implicit biases produce discriminatory behavior in persons who do not know that they have discriminatory biases. The best anecdotal evidence for lack of awareness of discriminatory implicit biases is the large proportion of people who, on self-testing with one or more of the freely available on-

line IATs, are surprised – often distressed – to learn that their test scores indicate more-than-trivial strengths of associations indicative of implicit bias.[9]

Resistance to change. Research showing that long-established implicit biases resist change has recently been reviewed in several authoritative publications. We describe here those reviews' findings and their significance. In 2009, psychologists Betsy Paluck and Donald Green reviewed a large collection of studies of prejudice reduction efforts and concluded that "Entire genres of prejudice reduction interventions, including moral education, organizational diversity training, advertising, and cultural competence in the health and law enforcement professions, have never been [rigorously] tested."[10] In 2021, Paluck, Green, and colleagues reported a follow-up review of several hundred subsequent studies, leading to their conclusion that "much research effort is theoretically and empirically ill-suited to provide actionable, evidence-based recommendations for reducing prejudice."[11] The discouraging conclusions of these two large reviews were preceded by a similarly discouraging 2006 review by psychologists Alexandra Kalev, Frank Dobbin, and Erin Kelly, who concluded that "Practices that target managerial bias through feedback (diversity evaluations) and education (diversity training) show virtually no effect in the aggregate."[12] Two substantial multi-investigator collaborative studies by psychologist Calvin Lai and colleagues, of experimental interventions designed to weaken or eliminate long-established implicit biases, concluded that these biases "remain steadfast in the face of efforts to change them."[13] That conclusion by Lai and colleagues was in striking contrast to the more optimistic conclusion – that automatic stereotypes and attitudes were "malleable" – from a 2002 review of the earliest studies of experimental interventions.[14] All of the interventions examined in the 2002 review had been tested with posttests administered very near in time to the intervention. In Calvin Lai, Allison L. Skinner, Erin Cooley, and colleagues' 2016 report of studies with 6,321 participants, none of eight interventions that had previously been found to be effective when tested near immediately after intervention was found to be effective in tests after delays ranging from several hours to several days.[15] The review articles we've briefly summarized here, along with others that reviewed studies conducted in other settings, have themselves been summarized more thoroughly in a recent review, which did not alter the overall picture.[16] We conclude that evidence for the effectiveness of methods assumed to be capable of reproducibly moderating or eliminating implicit biases is lacking.

With these properties of implicit biases in mind, we outline four misunderstandings of scientific work on implicit bias, each immediately followed by its evidence-based correction ("proper understanding").[17]

Misunderstanding 1: *IAT measures assess prejudice and racism*. Proper understanding: IAT measures reveal associative knowledge about groups, not hostility toward them. The IAT and other indirect measures are better described as measuring *bi-*

ases, a term that does not imply prejudice, hostility, or intent to harm, all of which are part of the generally understood meanings of "prejudice" and "racism."

Misunderstanding 2: *Implicit measures are capable of predicting only automatic behavior that is done unthinkingly. They do not predict intentional behavior that is done deliberately*. This misunderstanding was sufficiently widespread that one can find it stated in multiple peer-reviewed psychological publications of the last twenty years. Proper understanding: As three independently conducted meta-analyses have demonstrated, IAT measures equally predict automatic (spontaneous) and intentional (deliberate) behavior.[18]

Misunderstanding 3: *Implicit biases are amenable to modification by experimental treatment interventions*. Proper understanding: As we described above, published experimental tests do not find that long-established implicit biases are reliably modifiable, let alone eradicable, by interventions. This misunderstanding resulted from early studies that examined only effects observable within minutes of administering a treatment intervention. The effects of interventions that produced those findings are now known not to be durable.[19]

Misunderstanding 4: *Group-administered antibias or diversity-training procedures can effectively manage problems that have been attributed to systemic or implicit bias*. Proper understanding: The most authoritative reviews of available studies have concluded that the evidence falls far short of justifying such claims.[20]

How much discriminatory adversity is caused by implicit and systemic biases? Looking at implicit biases first, consider that majorities of all samples that have been studied display the race attitude IAT's "automatic White preference" result. Likewise, majorities (often including majorities of women) associate men more than women with career and women more with family, men more with leader roles and women more with support roles, and men more with STEM disciplines and women more with arts or humanities disciplines. In educational and work settings, these implicit biases predispose teachers and managers to judge the work of White persons more favorably than that of Black persons, to judge men more capable of leadership than women, and to judge men superior to women in math and science disciplines. These observations are a small portion of the empirical support for a conclusion that discrimination-predisposing implicit biases are present in majorities of most populations and, therefore, when aggregated over all those affected, must account for much more damage than do openly expressed (explicit) biases, which are never evident in more than small-to-modest minorities of research samples.

For systemic biases, consider that these are usually the result of a widely applied procedure that was (perhaps long past) created to serve organizational or governmental purposes, presumably without considering how it might affect demographic groups differentially.[21] Systemic biases occasionally receive attention

from public health organizations and news media. During 2021 and 2022, there was frequent reporting of racial disparities in health care outcomes for COVID-19, with substantial attention also given to disparities for groups differing in socioeconomic status or age. These disparities are sometimes striking enough to be perceived as unfair and to generate protest, but even so, those who notice the disparities are often in no position to either modify them or influence others to take corrective action. Many discriminatory systemic biases that have not been noticed sufficiently to generate public protest will continue to occur – implemented routinely by myriad employees of governments, businesses, hospitals, schools, and other institutions who are only doing the work that they were hired, elected, or appointed to do. Systemic biases can appear in the form of policies, practices, regulations, and traditions that typically affect multiple (often many) people and frequently produce relatively small effects – but their small size does not mean that those effects are ignorable. The small effects to individuals accumulate, both because of the large number of people affected and because they can affect the same persons repeatedly in settings such as work, school, shopping, travel, paying rent, and paying interest on loans.

There is presently no way to estimate with precision either the percentage of the U.S. population affected by discriminatory implicit and systemic biases, or the magnitude of adversity produced by those discriminatory impacts. We expect it to be relatively modest at the level of individual episodes. Even so, the number affected must be vastly greater than the very small percentage of the U.S. population that now seeks or obtains legal or other governmental redress for discrimination. We know this partly from studies of the Equal Employment Opportunity Commission (EEOC) by the Center for Public Integrity, which investigated the dispositions of discrimination complaints submitted to the EEOC from fiscal years 2010 through 2017.[22] The Economics Policy Institute has a report that goes beyond examining just the EEOC's actions, considering also its problems in gaining congressional budgetary support.[23] One cannot avoid concluding that a great deal slips through large cracks in governmental programs for dealing with discrimination in the United States, even if one considers only discrimination occurring in employment. It is certainly much greater than what is described in reports by the EEOC and parallel state-based agencies. And this is in a system that presently does not yet attempt to deal with more than a small fraction of discriminatory impacts of implicit and systemic biases. We gave brief thought to generating hypothetical estimates of costs, both to those who suffer discrimination and to organizations that have responsibilities for remedying discrimination. However, we are so far from having access to data that could allow even approximate estimates that we must let that challenge await later efforts. When economists with appropriate expertise do undertake such an accounting, they will not find that task easy. Damages due to implicit and systemic biases typically leave no fingerprints, let alone dollar signs.

Discrimination occurs in multiple domains that receive little attention from legislators, regulators, and courts. We learn about these the same way others do: from news reporting via a variety of media. In health care, differential diagnosis and treatment of persons of color, elderly persons, and impoverished persons have been documented in data and reporting from the Centers for Disease Control and Prevention (CDC), the Department of Health and Human Services (HHS), academic researchers, and investigative reporters. In real estate, properties belonging to racial and ethnic minorities are typically undervalued by realtors, meaning that owners receive artificially low offers when selling their properties. Minority purchasers are also most likely to be shown available rentals and homes selectively in neighborhoods in which their ethnic groups have an established presence, if not a majority. In banking, loans to African Americans, women, and members of other protected classes are more likely to be denied and loan interest rates are likely to be elevated. In insurance, as in banking and real estate, members of protected classes receive inferior service and coverage, higher rates charged, and lower rates of success of claims made by them as policy holders. In policing, there are thousands of daily interactions between law enforcement and African Americans and other members of protected ethnic and racial classes that produce increased stops, arrests, arraignments, injuries, and deaths.

Most people (we include ourselves) remain unaware of the majority of discrimination occurring around them. When workers suspect that they are being discriminated against, they will often have difficulty convincing coworkers, or even friends and relatives, that it is indeed discrimination. Should they file a discrimination complaint with the EEOC or other agency, those agencies are very often poorly funded or subject to the enforcement (or nonenforcement) interests of the political party currently in power. In some cases, the EEOC or other agency will investigate the claim and, if the process of conciliation with the accused is unsuccessful, file a lawsuit directly. But even then, agencies litigate a small portion of those lawsuits. More often, agencies will leave it to individual claimants to pursue a lawsuit themselves. Once a complainant receives a notice of right to sue from these agencies, they may, with the help of an attorney, pursue the case in court. Finding a lawyer who understands the claim adequately or finding funds to pursue the claim poses another series of barriers. An expert on implicit bias may also be needed to convince a judge that the plaintiff's case is one for which a jury might award damages. Before a trial occurs, the plaintiff who has overcome all these obstacles must often also survive a defendant's request for summary judgment that can lead a judge to decide to end the proceedings immediately in favor of the defendant.

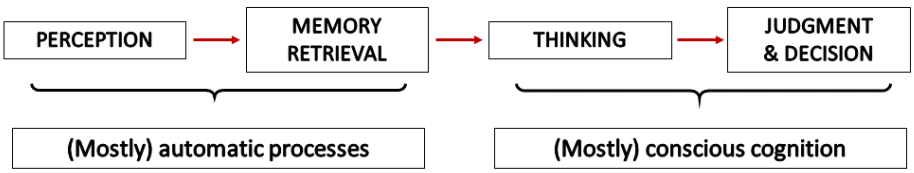
A suit strong enough to clear all these hurdles may lead the employer being sued to settle rather than face the probability of a jury finding for the plaintiff. Or uncertainty about a favorable outcome may prompt the plaintiff to accept a low

settlement offer. If a trial proceeds, something that happens for only a very small minority of cases, there still remains the barrier of rulings by the court on admissibility or sufficiency of evidence. In the end, a jury composed of persons who might not include a member of the plaintiff's protected class – the jury itself conceivably influenced by implicit biases – may reach a verdict against the plaintiff.

How do implicit biases produce discriminatory behavior? As we summarized above, when mental associations about demographic groups are triggered automatically, the associated attitudes and stereotypes influence behavior that may be discriminatory against members of those groups. But how do *the courts* consider implicit bias as a basis for unlawful discrimination? Because courts give close attention to the role of intent in contemporary discrimination law, and because "intent" is used with a variety of meanings in jurisprudence, we apply a definition of "intent to discriminate" based on the legal definition of intent provided in *Black's Law Dictionary*: intent to discriminate is the mental resolution or determination to do an act that existing law classifies as discriminating against a member of a protected class.[24] Using this definition, we conclude that implicit biases influence decisions that may prove discriminatory, even when decision-makers cannot be fully aware of this influence and may not anticipate that their actions will produce discriminatory consequences. The scientific basis for this understanding comes from adaptation of a long-established *information-processing stages* analysis of choice decision-making, as shown in Figure 2.

The processes of perception and memory retrieval in the first two stages of Figure 2 are understood to occur automatically when encountering a person.[25] Upon perception of stimulus features adequate to distinguish a demographic category of an encountered person (for example, their race), long-established associations, including ones measurable via the IAT, are activated in memory. This activation can predispose (or *prime* in psychology terms) conscious thoughts in the third (thinking) stage, and this priming can further influence conscious judgments and decisions made about the person in the fourth stage. These intentional fourth-stage decisions produce behaviors that can have discriminatory beneficial or harmful consequences, without the decision-maker being aware that these influences are acting through conscious thoughts and judgments that have been influenced by automatically activated mental associations (implicit biases). Psychologists describe the influence of the first two (automatic) stages on thought content as a *bottom-up* influence, meaning that lower (more rapidly occurring) mental processes are influencing higher (later occurring) processes. The influences of thought on judgment (third stage) and decision (fourth stage) are the influences of higher mental processes on behavior (that is, *top-down*). Another useful description of the mechanics of the model in Figure 2 is that the operations of the first two stages occur outside of conscious awareness, but the products of those

Figure 2
Information-Processing Stages Theoretical Framework for Choice Decision-Making



Source: Authors' image, as influenced by Edward E. Smith, "Choice Reaction Time: An Analysis of the Major Theoretical Positions," *Psychological Bulletin* 69 (2) (1968): 77.

operations are known consciously as they shape judgment and behavior in the third and fourth stages.

Any person with whom one interacts belongs to multiple demographic categories on dimensions of gender, race, ethnicity, weight, occupation, and socioeconomic status, among others. All familiar demographic categories have multiple associations that have been strengthened by overlearning since early childhood. It is therefore not surprising that an IAT measure of an association of just one demographic category with just one associated attitude or stereotype typically has only small-to-moderate correlations with measures of discriminatory attitudes or actions.

Two widely known Supreme Court decisions, *Brown v. Board of Education* in 1954 and *Price Waterhouse v. Hopkins* in 1989, featured critical evidence involving what is now understood as implicit bias.[26] These cases well preceded the scientific introduction of implicit bias in 1995.[27] The decision in *Brown* was influenced by findings of an experiment showing that, when given the choice between playing with a White or a Black doll, young Black schoolchildren were much more likely to choose the White doll. Their choices were implicit (indirect) expressions of a racial bias, because the bias in favor of white skin color was an indirect expression of the bias. In *Price Waterhouse*, the female plaintiff was turned down for promotion to a high position for which she was well-qualified. The decision-making executives at Price Waterhouse gave the explanation that her assertive personality, something they regarded as appropriate for a male occupant of the position for which she was being considered, was inappropriate as a trait of the

female plaintiff. This reaction to Hopkins's not conforming to the female gender stereotype of nurturance was an implicit (indirect) indicator that this gender stereotype had played a role in the firm's decision not to promote her.

Another widely known Supreme Court case, *Batson v. Kentucky* in 1986, concluded that peremptory dismissal of Black jurors solely on the basis of race constituted an equal protection violation. In a concurring opinion, Justice Thurgood Marshall wrote, "A prosecutor's own conscious or unconscious racism may lead him easily to the conclusion that a prospective black juror is 'sullen,' or 'distant,' a characterization that would not have come to his mind if a white juror had acted identically. A judge's own conscious or unconscious racism may lead him to accept such an explanation as well supported." Marshall's two examples of conscious or unconscious bias fit quite closely to our analysis, using Figure 2, of how implicit biases can influence judgments and decisions.

Recent scientific advances have led to a new understanding of how, when, and where discrimination occurs. Although scientific knowledge never achieves certainty, it does reach a point at which there is consensus. Scientific understanding of implicit bias is either at or close to that point in regard to the four established properties of implicit bias (pervasiveness, predictive validity of IAT, lack of awareness, and resistance to change). No one expects the American legal system to rapidly and efficiently accommodate new scientific understanding. Many proponents of change to the legal system would still say that change should not be rapid. It is (fortunately, many might say) not up to scientists to decide when new scientific understanding has developed to a point at which it should be put to use. In the world of federal jurisprudence in the United States, the Supreme Court primarily has that power, aided by the circuit Courts of Appeals.

Perhaps the best that scientists can do to advance a goal of change is to point out areas in which established science is at odds with current legal precedents in discrimination law.[28] In discrimination law, we see four areas of such discrepancy that might eventually prompt changes in law or jurisprudence.

First, present scientific understanding does not fit with the present difference in court precedents for what constitutes discrimination in employment law versus equal protection law. Both bodies of law have a requirement of intent; plaintiffs must demonstrate that the defendant intentionally discriminated. In employment law (Title VII), the intent requirement translates to the proposition that the defendant committed an action that caused adversity to the plaintiff under circumstances indicating that the plaintiff's protected class status was a causal factor. When the cause is implicit bias, the defendant may not understand the adversity-producing action as discriminatory – perhaps considering it appropriate, given the defendant's implicitly biased judgment of the plaintiff's job performance. In equal protection law (based primarily on the Fourteenth Amendment's declara-

tion that "No state shall . . . deny equal protection of the laws to any person within its jurisdiction"), the intent requirement translates to the proposition that the defendant did the action purposefully to cause harm to a member (or members) of the plaintiff's protected class. This purposeful intent requirement in equal protection law creates a high bar that reduces the likelihood of plaintiffs succeeding in equal protection cases, such as when newly legislated voting procedures impair the voting opportunities of members of a protected class. There is no apparent basis in scientific understanding of discrimination's mental underpinnings for this difference between evidence requirements of employment law and equal protection law.

A second science-based concern is the often-insurmountable requirement to demonstrate purposeful intent in equal protection cases. The legislators and other officials who create laws and regulations that may have been shaped by implicit or explicit biases may not have purposefully intended to create the resulting adversities, or they may have been careful not to leave evidence of purposeful intent. Violations of equal protection resulting from many governmental actions may therefore not have a path to redress in courts. Similarly, adversities resulting from systemic biases may only rarely exceed the purposeful intent requirement in the equal protection domain. It is difficult to understand why, for example, a state's discriminatory redistricting legislation that denies Black Americans proportional representation in legislative bodies should be treated as an equal protection violation only if plaintiffs can show that the enacting legislators were purposefully trying to reduce Black Americans' opportunities to vote.

A third science-based concern is the recent shift away from the use of *disparate impact* (and toward *disparate treatment*) as the legal criterion for identifying discrimination in employment law. Disparate impact is "The adverse effect of a facially neutral practice (esp. an employment practice) that nonetheless discriminates against persons because of their race, sex, national origin, age, or disability and that is not justified by business necessity. Discriminatory intent is irrelevant in a disparate-impact claim." Disparate treatment is "The practice, esp. in employment, of intentionally dealing with persons differently because of their race, sex, national origin, age, or disability. To succeed on a disparate-treatment claim, the plaintiff must prove that the defendant acted with discriminatory intent or motive."[29] Disparate impact (for which intent is not required) has long been regarded as the appropriate criterion to use when plaintiffs in employment suits claim discrimination due to a "pattern or practice" of the defendant (this translates to systemic bias, as used in this essay). Because of the need to demonstrate the defendant's intent when the court requires the disparate treatment criterion, those suits are necessarily more difficult for plaintiffs than are suits heard under the disparate impact requirements. As shown in a 2011 article by psychologist Lauren B. Edelman and colleagues, "disparate treatment has become far more

prevalent in civil rights cases over time," increasing from about 15 percent of cases in federal District Courts in 1970 to about 95 percent in 1997.[30] This was happening coincidentally with scholarly literature being on the verge of showing a decline in focus on intentional discrimination (see Figure 1). Courts' increasing focus on disparate treatment (for which evidence of intent is required) is at odds with recent social scientific and epidemiological work revealing the widespread operation of implicit and systemic biases, which can produce discrimination without accompanying evidence of intent to discriminate against members of protected classes. This would not happen if discrimination were, in the law, identified as behavior that causes adversity to protected classes rather than being identified with a state of mind that might (or might not) cause such adversity.

The fourth science-based concern is that, in employment discrimination class actions, implicit bias is not now recognized as a basis for establishing the existence of commonality, which is a requirement for certification of a class of plaintiffs in a discrimination suit. In the Federal Rules of Civil Procedure, the commonality requirement serves to assure that plaintiffs grouped into a class share the same basis for complaint against the employer.[31] To scientists, the pervasiveness of implicit biases seems a plausible and appropriate basis for commonality, but no plaintiff has yet tested this reasoning in a U.S. court.

How to deal with the great amount of discrimination that continues to occur in employment? The specifications of Titles VI and VII of the Civil Rights Act of 1964, including modifications added in subsequent congressional amendments and in Supreme Court and circuit Courts of Appeals precedents, fall well short of covering what scholarly and scientific work now identify as sources of employment discrimination. It is not simply the noncoverage of discriminatory impacts resulting from implicit and systemic biases. It is also that the Equal Employment Opportunity Commission's capacity does not come close to the EEOC's goals as stated in the Equal Employment Opportunity Act of 1972. The text of that Act starts with "The Commission is empowered . . . to prevent any person from engaging in any unlawful employment practice as set forth in section 703 or 704." Sections 703 and 704 contain the main statements of unlawful employment practices in the 1964 law's centerpiece, Title VII.

Writing this essay gave us some optimism that the science of implicit bias may be leverageable to improve prospects for plaintiffs to base effective discrimination suits at least partly on implicit bias evidence. Despite making good progress on that goal, much of what we learned in the process prompted us to consider prospects for effective efforts to address problems of discrimination outside the justice system, including both private-sector executives and officials in public-sector executive roles. We start with a short list of problems that can be addressed by actors in these nonjudicial, nonlegislative roles.

Efforts intended to remediate suspected or claimed discrimination in large organizations presently use training methods that are not established as effective. If they serve the organization at all, these training efforts do so by projecting the appearance that the organization's leaders are trying to eliminate or control discrimination. This almost always misleading (as it turns out) appearance can be counterproductive when it deflects leaders from seeking more effective methods.

Those who discover evidence of discrimination rarely occupy positions that enable them to work cooperatively with leaders of the organization in which they have uncovered discrimination. They are more likely to be seen as whistle-blowing enemies of the organization, possibly also becoming targets for retaliation. CEOs of large organizations may have little internal motivation and little external pressure to scrutinize the organization's personnel databases to identify discriminatory disparities that would be both easy to identify and straightforward to repair, once identified.

Many organizations assign responsibility for dealing with discrimination not to top-level leaders, but to organizationally subordinate human relations and legal departments, the personnel of which may have greater motivation to please their supervisors than to rock the organizational boat by investigating, discovering, and calling for remediation of discrimination within the organization.

We did not initially intend for this essay to propose private-sector remedies for discriminatory disparities due to implicit and systemic biases. That plan developed when we became aware of an underused remedial strategy, disparity-finding, that has three attractive properties: 1) it is easy to describe, 2) it is straightforward to administer, and 3) it can be deployed outside the American justice system.

Even though not previously named, the disparity-finding method is well known to epidemiologists, who use it frequently to find and identify public health problems.[32] These discoveries not only reveal health care disparities, but can also make apparent who is in the best position to fix the disparities. Consider this example: An epidemiologist working at Institute I discovers a health care disparity at Hospital H, where members of Group A are noticeably more likely to suffer from affliction X than are members of Group B. Alas, the researchers at Institute I may have no power to direct administrators or staff at Hospital H to undertake feasible remedies. For example, epidemiologists working for the CDC and for other research agencies uncovered numerous health care disparities during the COVID-19 pandemic. However, the CDC could not direct public health agencies or governmental officials in various localities to invest funds or otherwise take steps needed to implement fixes, even though it was often obvious what fixes would be required. This example makes clear why it is optimal for the work of disparity-finding to be the responsibility of executive personnel within the organization in which the disparity exists.

One often reads about investigative journalists uncovering discrimination, especially police profiling that clearly amounts to racial or ethnic bias. The journalists who reveal these problems are not in positions that enable them to implement fixes. That is the general problem in many contexts: the people with data capable of revealing the problem lack authority to intervene to fix the problem. Remarkably, this problem need not exist in many situations in which implicit biases or systemic biases are causing discriminatory disparities. In a business organization, the personnel data are owned by the company that employs the affected workers. In a police department, the data on racial characteristics of drivers and pedestrians stopped and searched by police officers, as well as the footage from body cameras operated by those officers, are in the possession of those police departments. In a university, records of qualifications and performances of students or staff who may be disadvantaged by implicit or systemic biases are in possession of the university itself. If the business organization, the police department, or the university employs a data scientist with appropriate quantitative skills, there should be no difficulty in using available data to uncover discriminatory disparities and report findings to administrative executives who can take responsibility for fixing them. How often does this sequence of disparity-finding followed by repair occur in organizations in which unrecognized discriminatory disparities exist? To the best of our knowledge – mainly because we almost never hear about it – the answer is "rarely."

All medium-to-large workplaces in the United States maintain personnel data as required by the EEOC and also as needed to keep their businesses operating. The available information usually includes employee demographics and data on employees' educational qualifications, productivity, job title, years employed, salary, raises, promotions, absences, performance evaluations, awards received, and discipline administered. If demographic disparities exist, the available personnel data likely contain evidence of them, and that evidence should not be difficult to find.

There is an essential second step after finding a disparity. A data analyst with the skills of an epidemiologist must also understand how interrelations among the personnel data variables can spuriously create or obscure appearances of a discriminatory disparity. Therefore, before suggesting that a discovered disparity must be repaired, a necessary step is to have the statistical expert assure that an identified disparity does not have a straightforward nondiscriminatory explanation. As just one easy to understand example, Group A might differ from Group B by having both 1) higher salaries and 2) stronger performance evaluations, even though the two groups are indistinguishable in qualifications, years employed, and other possibly relevant variables. This might be a basis for judging that Group A's greater average salary is explained by their superior performance and is therefore not discriminatory. However, that conclusion should await examining other possibilities, especially whether the performance evaluations were made objectively by a validated method or, instead, subjectively by the same manager who

determined each employee's pay. If the latter, the more plausible interpretation may be that the manager is discriminating in favor of members of Group A. For this reason, evaluations of performance are ideally done using objective criteria (that is, no subjective evaluation involved) by persons who play no role in deciding on pay or promotion.

There are reasons to believe that discriminatory disparities will be found almost whenever disparity-finding is undertaken. The two settings that produce most of the publicly known examples of disparity-finding occur in policing and health care, and that disparity-finding has been done mostly by outside agencies. In policing, watchdog/citizens' organizations and investigative reporters use FOIA (Freedom of Information Act) requests to obtain data access. In health care, the data may be voluntarily provided by hospitals or other medical institutions, or available in public archives such as those maintained by HHS or the CDC. Unfortunately, those efforts may not have enough data access to establish whether revealed disparities are discriminatory or nondiscriminatory. Also, an outside agency that obtains the information generally has no authority to grant, force, or enforce an effective fix of a discovered disparity. On the other hand, when disparity-finding is done within an organization that maintains its own personnel database, the finding is in the hands of those best positioned both to identify a plausible nondiscriminatory cause and to devise a fix if they cannot identify a nondiscriminatory explanation.

Business considerations, especially the standard goal of maximizing profit, may suppress willingness of organizational leaders to undertake routine (such as annual) disparity-finding scrutiny of their personnel data. Ideally, the CEO assigns responsibility for disparity-finding work to an executive whose annual bonus will increase directly as a function of success in identifying previously unrecognized disparities and determining whether they are discriminatory.

Advocates of internal disparity-finding should be aware that the organization's leaders will be concerned (appropriately) that employees who learn of uncovered disparities may use that knowledge to launch discrimination suits. For that concern not to discourage businesses from undertaking disparity-finding, courts can recognize a "self-critical analysis" privilege that protects the company from having its self-discovered evidence used against it. In practice, however, courts rarely grant this privilege, in effect motivating employers to neglect routine disparity-seeking scrutiny of their personnel data. Legal scholar Deana Pollard Sacks and others have pointed out that, if courts allow this self-critical analysis privilege, this could be very helpful in reducing unwanted discrimination, such as can result from not yet recognized implicit and systemic biases.[33]

We imagine how future historians may view progress of American treatment of discrimination since the Civil Rights Act of 1964. They might see that Act itself as a central piece in two centuries of legislation that

increased legal protections of civil rights beyond the Fifth Amendment's (1791) declaration that "No person shall be . . . deprived of life, liberty, or property, without due process of law" and the Fourteenth Amendment's (1868) assertion that "No state shall . . . deny equal protection of the laws to any person within its jurisdiction." Some important later pieces of legislative progress include the Fifteenth (1870), Nineteenth (1920), and Twenty-Fourth (1964) constitutional amendments, the Equal Pay Act (1963), the Age Discrimination in Employment Act (1967), and the Americans with Disabilities Act (1990). Concurrent with legislative developments since the middle of the twentieth century, decisions of the U.S. Supreme Court gradually limited the scope of antidiscrimination laws. Concurrently, but outside the legal system, scientists and scholars were establishing that much more discrimination than was previously apparent to the legal system was occurring in forms that were often not intended to harm and that were not readily apparent either to their perpetrators or to their victims. Remedy for those forms of discrimination – implicit and systemic biases – was not then easily available within the U.S. justice system.

Can we predict the next few sentences of this future history? An even more interesting question: what might be done now to shape the content of those sentences?

AUTHORS' NOTE

The authors' work in preparing this essay was aided substantially by multiple colleagues named here (in alphabetical order): Jan De Houwer, Alice H. Eagly, Lauren Edelman, Bertram Gawronski, Rachel Godsil, Cheryl Kaiser, Ian Kalmanowitz, Jerry Kang, Linda Hamilton Krieger, Calvin K. Lai, Robert S. Mantell, Brian Nosek, Rebecca G. Pontikes, Kate Ratliff, and James Sacher.

ABOUT THE AUTHORS

Anthony G. (Tony) Greenwald, a Fellow of the American Academy since 2007, is Professor Emeritus of Psychology at the University of Washington. He provoked modern attention to the psychological self with his 1980 article, "The Totalitarian Ego." His 1990s methods contributions made unconscious cognition and subliminal perception orderly research topics. His 1995 creation, the Implicit Association Test, enabled observation of unconscious associative knowledge, revamping understanding of stereotypes and prejudice.

Thomas Newkirk is a Partner at Newkirk Zwagerman. For thirty-five years, he has represented individuals who have been denied equal employment opportunities. He specializes in incorporating the science of implicit bias into the practice of law to advance the goal of achieving equality for all protected groups.

ENDNOTES

[1] "Implicit bias" was first used in its present meaning by Anthony G. Greenwald and Mahzarin R. Banaji in "Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes," *Psychological Review* 102 (1) (1995): 4–27.

[2] Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz, "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test," *Journal of Personality and Social Psychology* 74 (6) (1998): 1464–1480. This article introduced the IAT, originally presenting it as a measure of association strengths, later understood also as a measure of implicit biases.

[3] Kirsten N. Morehouse and Mahzarin R. Banaji, "The Science of Implicit Race Bias: Evidence from the Implicit Association Test," *Dædalus* 153 (1) (Winter 2024): 21–50, https://www.amacad.org/publication/science-implicit-race-bias-evidence-implicit-association-test.

[4] Ibid., 28.

[5] Ibid.

[6] Project Implicit, https://www.projectimplicit.net (accessed December 15, 2023).

[7] An overview and comparison of the three meta-analyses is available in Anthony G. Greenwald and Calvin K. Lai, "Implicit Social Cognition," *Annual Review of Psychology* 71 (2020): 419–445.

[8] The problem is that true experiments require assigning research subjects at random, early in life, to environments that would create the attitude and stereotype associations that comprise implicit biases. This is neither practical nor ethical.

[9] In *Blindspot: Hidden Biases of Good People*, Mahzarin R. Banaji and Anthony G. Greenwald reported that 40 percent of 1.5 million White Americans who had completed the Race attitude IAT online displayed the combination of explicit egalitarianism (self-reported equal liking of racial Black and racial White) together with IAT results indicating implicit preference for White relative to Black. Mahzarin R. Banaji and Anthony G. Greenwald, *Blindspot: Hidden Biases of Good People* (New York: Delacorte Press, 2013), 158.

[10] Elizabeth Levy Paluck and Donald P. Green, "Prejudice Reduction: What Works? A Review and Assessment of Research and Practice," *Annual Review of Psychology* 60 (2009): 356.

[11] Elizabeth Levy Paluck, Roni Porat, Chelsey S. Clark, and Donald P. Green, "Prejudice Reduction: Progress and Challenges," *Annual Review of Psychology* 72 (2021): 533–560.

[12] Alexander Kalev, Frank Dobbin, and Erin Kelly, "Best Practices or Best Guesses? Assessing the Efficacy of Corporate Affirmative Action and Diversity Policies," *American Sociological Review* 71 (4) (2006): 611.

[13] Calvin K. Lai, Maddalena Marini, Steven A. Lehr, et al., "Reducing Implicit Racial Preferences: I. A Comparative Investigation of 17 Interventions," *Journal of Experimental Psychology: General* 143 (4) (2014): 1765; and Calvin K. Lai, Allison L. Skinner, Erin Cooley, et al., "Reducing Implicit Racial Preferences: II. Intervention Effectiveness across Time," *Journal of Experimental Psychology: General* 145 (8) (2016): 1014.

[14] Irene V. Blair, "The Malleability of Automatic Stereotypes and Prejudice," *Personality and Social Psychology Bulletin* 6 (3) (2002): 242–261.

[15] Lai, Skinner, Cooley, et al., "Reducing Implicit Racial Preferences."

[16] Anthony G. Greenwald, Nilanjana Dasgupta, John F. Dovidio, et al., "Implicit Bias Remedies: Treating Discriminatory Bias as a Public Health Problem," *Psychological Science in the Public Interest* 23 (1) (2022): 7–40.

[17] A more detailed presentation of these misunderstandings is in Greenwald, Dasgupta, Dovidio, et al., "Implicit Bias Remedies."

[18] To illustrate, in the first of the three meta-analyses reported by Anthony G. Greenwald, T. Andrew Poehlman, Eric L. Uhllmann, and Mahzarin R. Banaji, each discriminatory bias measure that was tested for prediction by an IAT measure was rated for the extent to which the behavior was difficult to control consciously (0 = maximally spontaneous; 10 = maximally controllable). Choice of whom to vote for in a presidential election was a measure that was judged easy to control, whereas nonverbal behaviors such as eye blinks, speech hesitations, or body orientation variations were judged as highly spontaneous. Anthony G. Greenwald, T. Andrew Poehlman, Eric L. Uhllmann, and Mahzarin R. Banaji, "Understanding and Using the Implicit Association Test: III. Meta-Analysis of Predictive Validity," *Journal of Personality and Social Psychology* 97 (1) (2009): 17.

[19] A detailed review of studies of interventions can be found in Greenwald, Dasgupta, Dovidio, et al., "Implicit Bias Remedies." That review identified only a few isolated studies that found a measurable intervention effect after the day on which the intervention was administered. The review describes those exceptional results but was unable to identify any component of those studies that appeared responsible for findings of apparent durability. This leaves a presently unsolved mystery.

[20] For an overview that includes summaries of the authoritative reviews mentioned in this paragraph, see Greenwald, Dasgupta, Dovidio, et al., "Implicit Bias Remedies."

[21] It is easily conceivable that many systemic racial biases were created deliberately with the expectation that they would produce race or gender disparities.

[22] Maryam Jameel and Joe Yerardi, "Despite Legal Protections, Most Workers Who Face Discrimination Are on Their Own," The Center for Public Integrity, February 28, 2019, https://publicintegrity.org/inequality-poverty-opportunity/workers-rights/workplace-inequities/injustice-at-work/workplace-discrimination-cases. This report states, "Though the law places the burden on employees to prove discriminatory intent or impact, when hard evidence of unequal treatment exists, it is often buried in personnel records only the employer can access." See also Maryam Jameel, "More and More Workplace Discrimination Cases Are Closed Before They're Even Investigated," The Center for Public Integrity, June 14, 2019, https://publicintegrity.org/inequality-poverty-opportunity/workers-rights/workplace-inequities/injustice-at-work/more-and-more-workplace-discrimination-cases-being-closed-before-theyre-even-investigated.

[23] Jenny R. Yang and Jane Liu, "Strengthening Accountability for Discrimination," Economy Policy Institute, January 19, 2021, https://www.epi.org/unequalpower/publications/strengthening-accountability-for-discrimination-confronting-fundamental-power-imbalances-in-the-employment-relationship.

[24] *Black's Law Dictionary*, 11th edition (St. Paul, Minn.: Thomson Reuters, 2019). This elaboration of *Black's* definition of intent applies it to the context of discrimination law. This definition neither asserts nor implies that motive to inflict harm is part of intent.

[25] This point was established in a widely referenced theoretical article published in 1977: Richard E. Nisbett and Timothy D. Wilson, "Telling More Than We Can Know: Verbal Reports on Mental Processes," *Psychological Review* 84 (3) (1977): 231–259.

[26] *Oliver Brown, et al. v. Board of Education of Topeka, et al.*, 347 U.S. 483 (1954); *Price Waterhouse v. Ann B. Hopkins*, 109 U.S. 1775 (1989).

[27] See Greenwald and Banaji, "Implicit Social Cognition."

[28] Established science is never unanimously accepted by scientists. An essential and valuable aspect of the scientific method is to question existing understanding. Science becomes problematic when scientists stop questioning their favorite theories, which may be their own theories.

[29] These definitions are from *Black's Law Dictionary*, 11th edition.

[30] Lauren B. Edelman, Linda H. Krieger, Scott R. Eliason, et al., "When Organizations Rule: Judicial Deference to Institutionalized Employment Structures," *American Journal of Sociology* 117 (3) (2011): 888–954.

[31] A prerequisite for class certification stated in Rule 23(a) of the Federal Rules of Civil Procedure is that "there are questions of law or fact common to the class."

[32] This epidemiological technique was only recently given the label "disparity finding." See Greenwald, Dasgupta, Dovidio, et al., "Implicit Bias Remedies," 25.

[33] Deana A. Pollard, "Unconscious Bias and Self-Critical Analysis: The Case for a Qualified Evidentiary Equal Employment Opportunity Privilege," *Washington Law Review* 74 (1999): 913–1031.

Little Things Matter a Lot: The Significance of Implicit Bias, Practically & Legally

Jerry Kang

Skeptics point out that measures of implicit bias can only weakly predict discrimination. And it is true that under current technologies, the degree of correlation between implicit bias (for example, as measured by the Implicit Association Test) and discriminatory judgment and behavior is small to moderate. In this essay, I argue that these little effects nevertheless matter a lot, in two different senses. First, in terms of practical significance, small burdens can accumulate over time to produce a large impact in a person's life. When these impacts are integrated not only over time but double integrated over large populations, these little things become even more practically significant. Second, in terms of legal significance, an upgraded model of discrimination that incorporates implicit bias has started to reshape antidiscrimination law. This transformation reflects a commitment to "behavioral realism": a belief that the law should reflect more accurate models of human thinking and behavior.

Implicit bias is a concept that has diffused rapidly throughout our culture. One reason for the fast uptake is that it's intuitively obvious. Even without formal training in psychology or neuroscience, we realize that we navigate the world with limited cognitive resources. When confronted with a flood of sensory stimuli, what else can we do but use mental shortcuts to streamline our processing of that information. By automatically classifying any object we encounter into a category, we take advantage of our prior knowledge of and experience with that category to guide our response. For instance, if we recognize and classify something as a chair, we know how to pull it out from the table and sit down without a second thought. It doesn't matter whether that chair looks like an antique, a barstool, or an office chair, we know what a "chair" is and what to do with it. But just as we do this with chairs, we do this with people. We immediately classify a person we meet into multiple social categories, based on age, gender, race, and role. Next, meanings associated with those categories are automatically activated and guide our interaction with that person. None of this is surprising.

https://doi.org/10.1162/daed_a_02055

What is surprising is the possibility that the meanings associated with categories might be "implicit." By implicit, I mean they are not readily subject to direct introspection. In other words, I cannot fully ascertain the meanings (that is, the attitudes and stereotypes) that I have associated with a social category by simply asking myself for an honest account. We only have partial insight into the numerous mental associations stored in our brains, which operate automatically. Even though it's humbling to recognize that we lack perfect, introspective insight, this too isn't exactly shocking. Every time a smell, song, or taste triggers a once-forgotten memory, we realize that traces of the past remain in our minds even if we cannot access them at will.

Finally, the recent rise of generative artificial intelligence (AI) has highlighted the computer science problem of "garbage in, garbage out." If we train a chatbot using biased content (the garbage in), we should not be surprised that the chatbot spews biased content (the garbage out). But why would the computing machinery in our brains magically avoid this pitfall? If our own neural networks are trained through deep immersion in a social, economic, political, and media reality configured by status hierarchy, role expectations, culturally specific designations of friend versus foe, and media stereotypes, why would our brains automatically reject that learning?

In sum, one reason the concept of "implicit bias" has become so popular, so quickly, is because it makes intuitive sense. If we are honest about our limitations as thinking machines, we should not be surprised to learn that implicit biases exist and can alter our judgments. Of course, intuitive common sense is often dead wrong, so it's important to check against the scientific evidence. Since other contributions to this issue of *Dædalus* already do so, I won't repeat that work in detail. It suffices to say that:

1. "Implicit bias" is a valid scientific construct.
2. Implicit bias can be measured indirectly through various instruments, including reaction time measures such as the well-known Implicit Association Test (IAT).[1]
3. Implicit bias is pervasive (generally favoring in-groups and those higher on a social hierarchy); related to but different from explicit bias (measured via self-reports); and generally larger in magnitude than explicit bias on socially sensitive topics such as race (and other social categories).[2]
4. Implicit bias predicts real-world judgment and behavior in a statistically significant way, but the effect size is small to moderate.

Numerous scientific questions remain unanswered, but outright denial of the existence of implicit bias is no longer tenable. What remains unclear is how much implicit bias matters in real-world conditions. Also uncertain are the best ways

to counter implicit bias and its consequences. The focus of this essay is to unpack what it means for implicit bias to have "small-to-moderate" effect sizes. I argue that these "little things" matter a lot, in two senses. First, in terms of practical significance, small burdens can accumulate over time to produce a large impact in a person's life. When these impacts are integrated not only over time but double integrated over large populations, little things don't seem so little after all. Second, in terms of legal significance, an upgraded model of discrimination based on better science, including implicit bias, has started to reshape antidiscrimination law. This happens when those who make and interpret law embrace "behavioral realism:" a belief that the law should reflect more accurate models of human thinking and behavior.

Does implicit bias have a real-world impact? More precisely, does some measure of implicit bias, produced by an instrument such as the IAT, predict real-world discrimination? For this discussion, I define "discrimination" narrowly as treating someone differently because of perceived membership in a social category, even though everyone agrees that the social category should not influence the specific decision or behavior at hand. To answer this question based on all available research (and not just cherry-picked examples), we rely on meta-analysis. A meta-analysis is an analysis of analyses. Imagine an open-source collaboration that stitches together individual snapshots taken by different photographers, using different cameras, at different times, into a single panoramic, composite picture. But instead of photos, we use academic studies. More specifically, a meta-analysis calculates a single number from all the conducted research in a domain: in this case, an "effect size" that estimates the strength of the relationship between implicit bias and intergroup discrimination.

To date, three major meta-analyses have been conducted on the predictive validity of implicit bias by researchers across the ideological spectrum.[3] All meta-analyses found statistically significant effect sizes, which this literature states in terms of Pearson's *r*, the correlation coefficient.[4] The three meta-analyses, which used slightly different datasets and methodologies, calculated statistically significant correlations ranging from .10 to .24. Averaged over all three meta-analyses, the correlation is .165.[5] By convention, this effect size is called "small-to-moderate." To say that these correlations are "statistically significant" means roughly that they are unlikely due to chance. But most savvy readers know that *statistical* significance says little about *practical* significance.

One standard way to gauge practical significance is to square the *r* value to get the "percentage of variance explained." On the simplifying assumption of uniform variability in the *r* values measured across the meta-analyses, we get the r^2 value of .027 (.165 × .165 = .027). In other words, implicit bias would explain 2.7 percent of the total variance (a statistical term of art) measured in the intergroup

behavior. Your immediate reaction, even if you can't recall the statistical definition of variance, might be that this seems like a small percentage. Perhaps it's too small an effect for us to care about.

Indeed, this precise objection has long been raised by skeptical academics and advocates. For example, in 2005, legal scholar Amy Wax and psychologist Philip Tetlock editorialized in *The Wall Street Journal* that "there is often no straightforward way to detect discrimination of any kind, let alone discrimination that is hidden from those doing the deciding."[6] In 2009, Tetlock and legal scholar Gregory Mitchell worried that implicit bias researchers were politicizing science and raised objections to believing that implicit bias caused discrimination in the real world.[7] In their 2015 meta-analysis, psychologist Frederick Oswald and colleagues (including Tetlock and Mitchell) lamented that "researchers still cannot reliably identify individuals or subgroups . . . who will or will not act positively, neutrally, or negatively toward members of any specific in-group or out-group."[8]

In the legal domain, consider also the dismissive attitude reflected in court opinions rejecting the expert testimony of psychologist Anthony G. Greenwald, who invented the Implicit Association Test:

- "The application of Dr. Greenwald's cognitive theory on stereotyping to the circumstances at the Y[MCA] is speculative, *without any scientific basis*."[9]
- "This sort of superficial analysis . . . is not expert material; it is *the say-so of an academic who assumes* that his general conclusions from the IAT would also apply to [the defendant]."[10]

These examples demonstrate that the question of practical significance indeed remains a live controversy. How then should we think about the problem of small effect sizes?

First, we should not assume that small, measured *r* values are necessarily worthless. Back in 1985, cognitive psychologist Robert Abelson made this point powerfully with a baseball analogy. He asked, "What percentage of the variance in athletic outcomes can be attributed to the skill of the players, as indexed by past performance records?"[11] In simpler terms, how much does a typical player's batting skill (measured by batting average) explain the percentage of variance in any single at bat? The answer turned out to be spectacularly low: approximately one-third of 1 percent, which is equivalent to an effect size of $r = .056$.[12] Recall that the effect size measured for implicit bias was almost three times larger at $r = .165$. Even if we compared players whose batting averages were two standard deviations above the mean, to those who were two standard deviations below (roughly a .320 hitter compared to a .220 hitter), for a single at bat, skill would explain only 1.3 percent of the variance, which is equivalent to $r = .113$. With two outs in the final inning and a player in scoring position, every manager would re-

place a .220 hitter scheduled to bat with a .320 pinch hitter if available. But this reveals that a small-to-moderate effect size of $r = .113$ is practically significant in the multimillion-dollar sport of professional baseball.

Second, even if any single instance of discrimination caused by implicit bias seems trivial (such as a misperception or less friendly body language), we must consider their accumulation over time. Abelson explained his surprising findings by pointing out that batting skill manifests over multiple at bats during an entire season. As psychologists David Funder and Daniel Ozer elaborate:

> The typical Major League baseball player has about 550 at bats in a season, and the consequences cumulate. This cumulation is enough, it seems, to drive the outcome that a team staffed with players who have .300 batting averages is likely on the way to the playoffs, and one staffed with players who have .200 batting averages is at risk of coming in last place. The salary difference between a .200 batter and a .300 batter is in the millions of dollars for good reason.[13]

All this should remind us of the phrase "death by a thousand paper cuts." To integrate all implicit bias–actuated harms over time, we need to know frequency (how many "cuts" per unit of time) and duration (what time period to measure from beginning to end). Depending on the question, duration can be years at a firm, in an industry, in a career, and indeed one's entire lifetime. And frequency is not just one critical judgment every few years when we apply for a job or promotion. Instead, it could be every social, economic, political, and professional interaction. It could be every time we get into a parking dispute; every time we get pulled over for a traffic stop; every time we ask for help at a hardware store; every time we shop for furniture, a car, or a house; every time we apply for a credit card or loan; every time we wait to be seated at a restaurant; every time we apply for a job or promotion; every time we turn in mediocre work and get (or don't get) the benefit of the doubt; every time we join a team; every time credit is shared; every holiday office party; and so on. In some sense, the frequency is multiple times per day because almost no social interaction is immune from implicit biases. This amounts to far more than a thousand cuts.

Third, after integrating "paper cuts" across time to assess an individual's harm, we should double integrate over all people potentially affected. An illuminating example comes from public health. Back in 1990, psychologist Robert Rosenthal pointed out that in a clinical trial, scientists noticed a statistically significant correlation of $r = .034$ between taking aspirin and reduced chances of heart attack.[14] Even though the correlation was small (almost five times smaller than the effect size for implicit bias), the scientists stopped the randomized double-blind study because they felt it was not ethical to continue giving the control group placebos. In Greenwald's account of that study, he considered the population-level impacts of decreasing the chances of heart attack even marginally for each participant.[15]

Given the millions of people subject to heart attack, aspirin could prevent approximately four hundred and twenty thousand heart attacks over a five-year period – something we all presumably agree is practically significant. Considering these three lessons – the batting average, a thousand paper cuts, and double integrals across time and people – will produce a more thoughtful understanding of practical significance. Still, having more concrete examples is helpful, and one way to produce them is to run simulations under plausible assumptions.

For example, Greenwald and psychologists Mahzarin R. Banaji and Brian Nosek modeled the potential impact of implicit bias on racial profiling. Suppose that implicit bias nudges police officers to cite Black drivers and pedestrians more frequently than White ones. Assuming that the effect size was just $r = .148$ (a value calculated in one of the Oswald meta-analyses highly critical of implicit bias), Greenwald and colleagues imagined two different worlds. In World 1, all the police officers were one standard deviation lower on implicit bias, and in World 2, all the police officers were one standard deviation higher on implicit bias. If we compared these two worlds, World 1 would have 9,976 fewer Black stops, which amounted to 5.7 percent of the total number of stops for the year of data analyzed.[16] Who would argue that avoiding nearly ten thousand police stops of Black people annually is practically insignificant?

In another example, Greenwald created a simulation to estimate how much implicit bias could alter the expected prison sentence for committing a crime. With plausible assumptions (a crime with a mean sentence of five years and a standard deviation of two years), implicit bias effect size of $r = .10$, and a five-round model (involving arrest, arraignment, plea bargain, trial, and sentencing), the simulation found that a Black criminal can expect a probabilistic sentence of 2.44 years versus a White criminal expecting 1.40 years. Remember that we must integrate this individual-level differential over the entire relevant population of criminal cases in any given year, which can run into the tens of thousands.[17] Even if there were only one thousand cases of this sort per year, implicit bias would produce one thousand years of more Black imprisonment annually. Again, how can this be practically insignificant?

Consider one last simulation involving Big Law. Assume that, to make partner, litigation associates must survive a monthly up-or-out tournament that lasts for eight years. Suppose that implicit bias creates just a 1 percent difference in the monthly survival rate, with the White associate likely to survive at 99 percent but the Asian associate likely to survive at 98 percent.[18] For simplicity's sake, if we assume each month's survival rate to be an independent probability, the White associate's chances of making partner (which requires surviving $8 \times 12 = 96$ cuts) would be 38.1 percent, whereas the chances for the Asian associate would be 14.4 percent.[19] And what is the r value equivalent for that 1 percent difference in

monthly survival rate? It amounts to a mere $r = .04$. The reason why such a small correlation can produce such drastic results is because each month a critical decision (up or out) is being made, and we are considering the accumulated impact of such decisions over ninety-six months.

There are bones to pick, of course, with the above simulations as being too stylized and not realistic. One can also object that the predictive validity studies were not conducted out in the field, under real-world circumstances, which include legal and procedural checks on discrimination. These are fair criticisms. But when we insist on greater realism, better evidence, or larger effect sizes, we should do so consistently, without double standards. For example, let's compare implicit bias to medical phenomena that we generally accept as practically significant. We already made one such comparison with aspirin and heart attacks. In 2001, psychologist Gregory Meyer and colleagues compiled a useful inventory of the effect sizes of what might be called medical "common sense."[20] Interestingly, they were often lower than or on par with the effect size found for implicit bias ($r = .165$):

- antihypertensive medication and reduced risk of stroke ($r = .03$),
- chemotherapy and surviving breast cancer ($r = .03$),
- antibiotic treatment of acute middle ear pain in children and improvement within seven days ($r = .08$),
- alcohol use during pregnancy and subsequent premature birth ($r = .09$),
- combat exposure in Vietnam and subsequent PTSD within eighteen years ($r = .11$),
- extent of low-level lead exposure and reduced childhood IQ ($r = .12$),
- nonsteroidal anti-inflammatory drugs and pain reduction ($r = .14$),
- post–high school grades and job performance ($r = .16$),
- validity of employment interviews for predicting job success ($r = .20$), and
- effect of alcohol on aggressive behavior ($r = .23$).

Given these comparable effect sizes, will those who object to the practical significance of implicit bias similarly object to the practical significance of these other phenomena? Second, let's compare the practical significance of implicit bias with that of explicit bias. When we discover that someone has explicit bias, we typically take note. For example, when meeting a new neighbor, if they blurt out anti-Semitic tropes, we will presumably take note. Similarly, during *voir dire* (the process of questioning potential jurors), if someone expresses stereotypes that Latinos are culturally prone to criminal gang activity, we will again take note. When we notice such expressions of explicit bias, we don't chastise ourselves for being irrational, credulous, "woke," or ideological. But here's where things get in-

teresting: the same meta-analyses that found small-to-moderate effect sizes for implicit bias revealed that implicit bias scores have comparable or *more* predictive power than explicit bias scores.[21] This suggests that if we take explicit bias seriously (because it might predict discriminatory judgment and behavior), we should take implicit bias even more seriously.

Third, let's compare the effect size of implicit bias with effect sizes that are often deemed legally significant in civil rights enforcement. The Equal Employment Opportunity Commission's (EEOC) Uniform Guidelines on Employee Selection Procedures adopt a rule of thumb that when a selection rate for any protected category is less than four-fifths of the rate for the group with the highest success rate, this disparity will be regarded as *prima facie* evidence of adverse impact, which is the first step of winning a disparate impact case under Title VII of the 1964 Civil Rights Act.[22] What does this rule of thumb mean in terms of effect sizes? Consider the following hypothetical about junior-level promotions in a national firm.

Among White applicants in any given year, suppose that five hundred are promoted and five hundred are not. In other words, the promotion rate for White people is 50 percent (five hundred out of one thousand total). Next, suppose that among Asian applicants (a smaller population), thirty-nine are promoted and sixty-one are not; the promotion rate is thus 39 percent (thirty-nine out of one hundred total). Because the ratio of promotion rates (39 percent Asian to 50 percent White) is lower than four-fifths, agency guidelines instruct judges to find *prima facie* evidence of a disparate impact. What do these differences in promotion rates look like when they are converted into Pearson's *r*? The $r = .063$. In other words, the federal government has announced a rule of thumb suggesting legal significance – under plausible assumptions of population size and promotion rates – for an effect size that is only $r = .063$. On what grounds, then, can we reflexively dismiss implicit bias ($r = .165$) as practically insignificant?

In sum, a careful inquiry into practical significance reveals that phenomena with small effect sizes can be practically significant. Little things mean a lot, not only in the trajectory of individual lives but also in the arc of entire peoples. In addition, we should actively scan for double standards. For example, if we happily rely on medical common sense – as we pop supplements, avoid heart attacks, or decide on treatment for breast cancer – we should recognize that we do so often because of *r* values *lower* than the effect sizes found with implicit bias. If we dismiss implicit bias as practically insignificant, then what justifies the double standard in our own self-care? Could it be that we worry about our own health and beauty but not so much about implicit bias – mediated harms inflicted on others?

Also, if we care so deeply about explicit bias, enough to interrogate potential jurors about their prejudices and stereotypes publicly during *voir dire*, on what sci-

entific grounds should we dismiss implicit bias as unimportant? To recap, over the past three decades in the mind sciences, researchers have uncovered surprising evidence that discrimination may be caused by implicit bias. How should these new discoveries influence the law? For two decades, I have advocated for a school of thought called "behavioral realism," which combines the traditions of legal realism and behavioral science. Stated succinctly, behavioral realism insists that law should incorporate more realistic models of human behavior.

This approach involves a three-step process.[23] First, we should regularly scan the sciences for more accurate, upgraded models of human decision-making and behavior. Second, we should compare that upgraded model to the "commonsense" legacy understandings embedded within the current law. Third, when the gap between the upgraded and legacy models grows sufficiently large (however defined), we should revise the law or its interpretation in accordance with the upgraded model. If that can't be done – for example, because of controlling precedent, constitutional constraints, or other overriding moral or policy considerations – then lawmakers should clearly explain their reasons why.[24] This requirement applies to judges and administrative agencies, in particular, who are obliged to give reasons for how they interpret and make law.[25] This simple three-step process largely avoids contentious normative questions and instead draws on a broadly overlapping consensus regarding 1) promoting instrumental rationality and 2) avoiding hypocrisy.

Concerning instrumental rationality, importing an upgraded, more behaviorally realistic model of decision-making means that the law will function under more accurate descriptions of human action. Doing so will be more efficient. For example, if the mind sciences discover better ways to deter bad behavior in adolescents, white-collar criminals, and large corporations, it would be instrumentally rational to incorporate these insights into our legal deterrence regimes. Concerning hypocrisy, all laws, including antidiscrimination laws, have some publicly announced purpose. When we learn that their purpose cannot be well-achieved because we are relying on legacy understandings, we should do something about it. If we decline to do so without good reason, we risk hypocrisy. For example, suppose a bank adopted cybersecurity measures – such as firewalls, multifactor authentication, and password managers – to prevent online fraud and other security breaches. But the bank discovers that its measures have failed all along because they fundamentally misunderstood underlying vulnerabilities like social engineering. If the bank declines to adapt to this realization, can we believe that it cares about security? And if it continues to tout its commitment to security, would we not criticize such advertising as deluded or hypocritical? As I elaborate below, this simple approach of behavioral realism has already started to influence antidiscrimination law.

A central feature of American civil rights law is the stylized distinction between intentional discrimination and disparate impact. On one hand, most antidiscrimination laws require a showing of intentional discrimination, which generally means that the defendant purposefully treated someone differently because of their social category. The focus is on the mental state of the individual defendant and their deliberate, purposeful consideration of a social category. On the other hand, some civil rights laws require only a showing of disparate impact.[26] As long as a specific practice causes a disparate impact across legally protected social categories, that practice must be specially justified.

In the employment context, a disparate impact–causing practice must be functionally necessary in the sense that it must be job-related and a business necessity. In addition, if there is an alternative policy or practice that produces equally good results with less disparate impact, the defendant will be held liable if they refuse to adopt it. The focus of disparate impact liability is not on the individual defendant's state of mind; instead, it is on group consequences. Even without legal training, one can see how disparate impact theory casts a broader net for legal concern than intentional discrimination. After all, many facially neutral selection criteria, adopted and applied without purposeful intentional discrimination, can produce a disparate impact.

For example, if there is an average height difference between Asian Americans and White Americans, then a minimum height requirement for first responders – originally adopted and applied without consideration of race – can produce a disparate racial impact. It was precisely this anxiety of disparate impact overreach that led the Supreme Court to read the federal Constitution's Equal Protection Clause narrowly, to proscribe only intentional discrimination. The historic case was *Washington v. Davis* (1976).[27] In that case, the question presented was whether a particular qualifying test that produced a disparate impact on Black police officer candidates violated their federal constitutional equal protection rights. The court explained that because there was no purposeful intent to harm Black candidates, there was no constitutional infirmity. The court's policy rationale was explicit:

> A rule that a statute designed to serve neutral ends is nevertheless invalid, absent compelling justification, if in practice it benefits or burdens one race more than another would be far-reaching and would raise serious questions about, and perhaps invalidate, a whole range of tax, welfare, public service, regulatory, and licensing statutes that may be more burdensome to the poor and to the average black then to the more affluent white.[28]

Intentional discrimination remains the constitutional touchstone and the initial presumption in interpreting all antidiscrimination laws. Moreover, as noted above, "intentional" is often presumed to mean "purposeful" and not a lower lev-

el of culpability, such as "knowing," "reckless," or "negligent."[29] Unfortunately, proving that the defendant purposefully treated someone worse because of a protected social category is extraordinarily difficult. That's why Critical Race Theorists have criticized the intentional discrimination requirement as privileging the "perpetrator perspective."[30] Has the science of implicit bias, by way of behavioral realism, weakened this fixation? Consider the following examples.

In *Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc.* (2015),[31] the Supreme Court had to interpret the federal Fair Housing Act (FHA), which declares it unlawful "to refuse to sell or rent, or otherwise make unavailable . . . a dwelling to any person because of race [and other protected categories]."[32] The question presented was whether the statute required purposeful intentional discrimination, or might it also recognize disparate impact? In a 5–4 decision, the court recognized a disparate impact theory of liability. Per Justice Anthony Kennedy:

> Recognition of disparate-impact liability under the FHA also plays a role in uncovering discriminatory intent: It permits plaintiffs to counteract *unconscious prejudices* and disguised animus *that escape easy classification* as disparate treatment. In this way disparate-impact liability may prevent segregated housing patterns that might otherwise result from covert and illicit stereotyping.[33]

According to the court, the more capacious disparate impact theory of liability was better suited to respond to "unconscious prejudices." Partly because the court accepted an upgraded model of human decision-making, which included the possibility of discrimination based on implicit social cognitions, the court adopted a broader interpretation of the Fair Housing Act to include disparate impact liability.

In *Kimble v. Wisconsin Department of Workforce Development* (2010), the Eastern District of Wisconsin heard an employment discrimination case under Title VII of the 1964 Civil Rights Act.[34] Title VII recognizes both disparate treatment (intentional discrimination) and disparate impact theories of liability. For disparate treatment, courts frequently suggest that the defendant must have explicitly and purposefully used a protected social category in its decision-making. But in truth, the statute does not specify any such mental state. Instead, it simply prohibits employment discrimination "because of" a person's race and other protected social categories. With this textual flexibility in mind, the court pivoted away from purposeful intent and instead asked more literally for category causation.[35] It explained "[n]or must a trier of fact decide whether a decision-maker acted *purposively*. . . . Rather, in determining whether an employer engaged in disparate treatment, the critical inquiry is *whether its decision was affected* by the employee's membership in a protected class."[36] Applying this clarified legal understanding to the facts of the

case, the court observed that "when the evaluation . . . is highly subjective, there is a risk that supervisors will make judgments based on stereotypes of which they *may or may not be entirely aware*."[37] It noted that because of the ordinary psychological process of categorical thinking, a supervisor may use stereotypes "*whether or not* the supervisor is *fully aware* that this is so."[38] Again, an upgraded model of discrimination, which the court gleaned in part from secondary sources advocating behavioral realism, led the court to rule in favor of the plaintiff.[39]

In *State v. Gill* (2019),[40] the Court of Appeals for Kansas had to interpret a state statute that prohibited "racial or other biased-based policing."[41] The case was prompted by a police officer approaching two Black men in an SUV because they were allegedly "staring hard" at him, which resulted in a search that uncovered drugs. The trial court found a statutory violation, and on appeal, the appellate court affirmed. The dissent railed loudly at the majority for "brand[ing] an officer of the law . . . a racist . . . . [without] evidence supporting such a serious charge."[42] But importing an upgraded, more behaviorally realistic model of discrimination, the majority de-escalated and explained that "no one here is branding [the officer] a racist."[43] Instead, the relevant question was one of racial causation, whether the officer "let racial bias – *conscious or unconscious – affect* his initiation of enforcement action."[44]

In *Woods v. City of Greensboro* (2017),[45] the Fourth Circuit Court of Appeals reviewed a district court's granting of a motion to dismiss a 42 U.S.C. § 1981 civil rights action (equal contracting rights) for failure to state a claim. The appellate court started its analysis by noting that "many studies have shown that most people harbor *implicit biases* and even well-intentioned people *unknowingly act* on racist attitudes."[46] Showing psychological sophistication, the court pointed out that the same actor may discriminate differently depending on the context: "it is unlikely today that an actor would *explicitly discriminate* under all conditions; it is much more likely that, where discrimination occurs, it does so in the context of *more nuanced decisions* that can be explained based upon reasons *other than illicit bias*, which though *perhaps implicit, is no less intentional*."[47]

Finally, the court warned that: "there is thus a real risk that legitimate discrimination claims, particularly claims based on more *subtle theories* of stereotyping or *implicit bias*, will be dismissed should a judge substitute his or her view of the likely reason for a particular action in place of the controlling plausibility standard."[48] For these reasons, the Court of Appeals reversed the dismissal and allowed the case to proceed to discovery.

The Supreme Court of the State of Washington deserves special recognition as a trailblazer for behavioral realism. Consider, for example, how it has evolved the processing of peremptory challenges. Way back in *Batson v. Kentucky* (1986), the United States Supreme Court held that a prosecutor's pur-

poseful discrimination to strike jurors because of race violated federal equal protection guarantees.[49] Unfortunately, it was nearly impossible to prove such a state of mind because any competent prosecutor could provide non-race-based justifications for striking a potential juror.

In 2018, the Washington Supreme Court pivoted away from demanding proof of a prosecutor's *subjective* mental state. Instead, the court adopted an *objective* reasonable person standard via judicial rulemaking (General Rule 37) and opinion in *Washington v. Jefferson* (2018).[50] Their revised approach asks whether "an *objective observer* could view race or ethnicity *as a factor* in the use of the peremptory challenge."[51] What's fascinating is that this objective observer benefits from a fully upgraded model of discrimination. General Rule 37(f) expressly states: "For purposes of [the Nature of Observer] rule, an objective observer is aware that *implicit, institutional, and unconscious biases*, in addition to purposeful discrimination, have resulted in the unfair exclusion of potential jurors in Washington State."[52]

Other states have followed Washington's lead. For example, in 2020, the California legislature passed AB 3070, which targeted "the use of group stereotypes and discrimination, whether based on *conscious or unconscious bias*, in the exercise of peremptory challenges."[53] California's statute does not require proof of intentional discrimination; instead, upon a challenge, the court must determine whether "there is a substantial likelihood that an *objectively reasonable person* would view race [and other protected categories] as *a factor* in the use of the peremptory challenge ...."[54]

In 2021, the Arizona Supreme Court eliminated all peremptory challenges in part due to the problem of implicit bias.[55] In 2022, upon the recommendations of a judicial task force, the judges of Connecticut's Superior Court amended their Practice Book not to require any showing of purposeful discrimination. Instead, courts must now ask whether the peremptory challenge "as *reasonably* viewed by an *objective* observer, legitimately raises the *appearance* that the prospective juror's race or ethnicity was a factor."[56] Similar to the State of Washington's approach, the objective observer "is aware that purposeful discrimination, and *implicit*, institutional, and *unconscious* biases, have historically resulted in the unfair exclusion of potential jurors."[57]

Finally, in 2022, New Jersey's Supreme Court amended its Rules Governing the Courts of the State of New Jersey to no longer require a showing of "purposeful discrimination." Instead, courts must now ask whether "a *reasonable, fully informed person* would view the contested peremptory challenge" to be based on a protected social category.[58] The Official Comment lists reasons that are presumptively invalid because they are historically associated with "improper discrimination, explicit bias, and *implicit* bias."[59]

Consider also how the Washington Supreme Court diverged from the path created by *McCleskey v. Kemp* (1987).[60] In *McCleskey*, the United States Supreme Court

declined to find an Eighth Amendment federal constitutional violation based on statistical evidence showing gross racial disparities in capital punishment. The court explained:

> At most, the [statistical] study indicates a discrepancy that appears to correlate with race. Apparent disparities in sentencing are an inevitable part of our criminal justice system. . . . Where the discretion that is fundamental to our criminal process is involved, we decline to assume that what is unexplained is invidious . . . . [W]e hold that the [statistical] study does not demonstrate a constitutionally significant risk of racial bias.[61]

Nearly three decades later, in *Washington v. Gregory* (2018), the Washington Supreme Court explained the importance of revising law "in light of 'advances in the scientific literature.'"[62] In its clearest endorsement of behavioral realism, the court explained: "where new, objective information is presented for consideration, we must account for it. Therefore, Gregory's constitutional claim must be examined in light of the newly available evidence presented before us."[63] The court then alloyed statistical evidence of racial disparities in capital punishment with an upgraded psychological model of discrimination to find a state constitutional violation:

> Given the evidence before this court and our judicial notice of *implicit* and overt racial bias against black defendants in this state, we are confident that the association between race and the death penalty is *not* attributed to random chance. We need *not* go on a fishing expedition to find evidence external to [the statistical] study as a means of validating the results. Our case law and history of racial discrimination provide ample support.[64]

Although statistics alone were not enough in 1987 for the federal Supreme Court, statistics coupled with general awareness of implicit bias sufficed for the state of Washington in 2018.[65] As the above examples demonstrate, by embracing behavioral realism, courts have imported more accurate models of discrimination that account for implicit bias. And through these upgraded understandings, courts have interpreted and applied both substantive and procedural laws differently than they would have under legacy beliefs. These cases evince the legal significance of implicit bias.

To be clear, these examples speak more to future potential than current actualization. As pointed out above, in the discussion of effect sizes, there are many courts that dismiss implicit bias as politicized, exaggerated, inflammatory, and too general to help decide specific cases. In addition, the censorship of so-called dangerous ideas, such as Critical Race Theory and implicit bias, will exact its political toll. But as also demonstrated above, we have already witnessed significant examples of legal transformation based on the evidence of implicit bias.

Intriguingly, the Supreme Court's recent, aggressive turn toward "but-for" causation in antidiscrimination law may spawn still more opportunity. In *Comcast Corporation v. National Association of African American-Owned Media, et al.* (2020), the Supreme Court adopted a baseline understanding based on "'textbook tort law' that . . . a plaintiff must demonstrate that, but for the defendant's unlawful conduct, its alleged injury would not have occurred. . . . That includes when it comes to federal anti-discrimination laws . . . ."[66] The court's objective was to restrict "mixed motive" cases, in which the plaintiff could prevail on a discrimination claim if race (or some other protected social category) was one "motivating factor" among many, even if it were not the "but-for" cause.

Consider the unexpected opportunity that this standard creates, however, for incorporating implicit bias.[67] If we take "but-for" causation seriously, that means we ask a simple counterfactual question: if the Black person were White, would they have been treated the same? We do not have to make findings about purposeful intent or whether the defendant subjectively and self-consciously considered race, which is so hard to prove. Instead, we are simply left with a probabilistic question of fact, about "but-for" causation, to be decided by the fact finder, based on all admissible evidence and their model of human decision-making.

The science of implicit bias is paradoxically both intuitive and disorienting. On the one hand, we know that our brain leverages schemas and categories to efficiently process the world, and the fact that we might do so with human social categories should not surprise us. On the other hand, because we have been taught that discrimination is wrong, it disorients us to find out that we may be discriminating without even realizing. A natural defensive reaction is to simply dispute the science as incorrect. When outright denial is impossible, given that the findings are statistically significant, the next step is to minimize the harm and deny their practical significance because of low effect sizes. As I have demonstrated, however, little things matter a lot. And if we resist double standards, we see that implicit bias is indeed a matter of practical significance for individuals and for society.

These new facts about implicit social cognition have provided us with a more behaviorally realistic model of discrimination. This upgraded model has rapidly diffused throughout our culture and has made inroads even into the staid law. It would be naive to assume that by virtue of greater accuracy and realism the model will necessarily prevail. Surely politics and ideologies will have their say. But over the past quarter-century, the evolving science of implicit bias has presented us with a stark choice. We can act like ostriches, burying our heads in the sand, and selectively insist on metaphysical certitude before taking corrective action. Or we can concede our cognitive limitations, roll up our sleeves, and try to design better policies, procedures, practices, and even laws to prevent discrimination from its various causes – including implicit bias.

ABOUT THE AUTHOR

Jerry Kang is Distinguished Professor of Law and Distinguished Professor of Asian American Studies (by courtesy) at the University of California, Los Angeles. He is the author of *Communications Law and Policy: Cases and Materials* (with Alan Butler and Blake E. Reid, 2023) and *Race, Rights, and Reparation: Law and the Japanese American Internment* (with Eric K. Yamamoto, Margaret Chon, Carol L. Izumi, and Frank H. Wu, 2013).

ENDNOTES

[1] See Anthony G. Greenwald and Calvin K. Lai, "Implicit Social Cognition," *Annual Review of Psychology* 71 (1) (2020): 419, 423–424, providing three categories of instruments: 1) the Implicit Association Test (IAT) and its variants; 2) priming tasks (where brief exposure to priming stimuli facilitates or inhibits subsequent reactions); and 3) miscellaneous other tasks including linguistic or writing exercises. For descriptions of the IAT, see Kate A. Ratliff and Colin Tucker Smith, "The Implicit Association Test," *Dædalus* 153 (1) (Winter 2024): 51–64, https://www.amacad.org/publication/implicit-association-test. In the legal literature, see Jerry Kang and Kristin Lane, "Seeing through Colorblindness: Implicit Bias and the Law," *UCLA Law Review* 58 (2) (2010): 465, 472–473; and Kristin A. Lane, Jerry Kang, and Mahzarin R. Banaji, "Implicit Social Cognition and the Law," *Annual Review of Law and Social Science* 3 (1) (2007): 427, 428–431.

[2] Kirsten N. Morehouse and Mahzarin R. Banaji, "The Science of Implicit Race Bias: Evidence from the Implicit Association Test," *Dædalus* 153 (1) (Winter 2024): 21–50, https://www.amacad.org/publication/science-implicit-race-bias-evidence-implicit-association-test; Kate A. Ratliff, Nicole Lofaro, Jennifer L. Howell, et al., "Documenting Bias from 2007–2015: Pervasiveness and Correlates of Implicit Attitudes and Stereotypes II" (unpublished pre-print); and Brian A. Nosek, Frederick L. Smyth, Jeffrey J. Hansen, et al., "Pervasiveness and Correlates of Implicit Attitudes and Stereotypes," *European Review of Social Psychology* 18 (1) (2007): 36–88.

[3] Anthony G. Greenwald, T. Andrew Poehlman, Eric Luis Uhlmann, and Mahzarin R. Banaji, "Understanding and Using the Implicit Association Test: III. Meta-Analysis of Predictive Validity," *Journal of Personality and Social Psychology* 97 (1) (2009): 17, 19–20, explaining that $r = .24$ for Black/White bias; Frederick Oswald, Gregory Mitchell, Hart Blanton, et al., "Predicting Ethnic and Racial Discrimination: A Meta-Analysis of IAT Criterion Studies," *Journal of Personality and Social Psychology* 105 (2) (2013): 171–192, explaining that $r = .15$ on Black/White implicit bias; and Benedek Kurdi, Allison E. Seitchik, Jordan R. Axt, et al., "Relationship Between the Implicit Association Test and Intergroup Behavior: A Meta-Analysis," *The American Psychologist* 74 (5) (2019): 569–586.

[4] The correlation coefficient indicates the strength of the linear relationship between two variables, in this case implicit bias and intergroup behavior. If the relationship is perfectly linear, then $r = \pm 1.0$, where a +1 value indicates a perfectly positive linear relationship and a –1 indicates a perfectly negative linear relationship. A value of $r = .0$ would indicate that there is no linear relationship between the two variables.

[5] Anthony G. Greenwald, Nilanjana Dasgupta, John F. Dovidio, et al., "Implicit-Bias Remedies: Treating Discriminatory Bias as a Public-Health Problem," *Psychological Science in the Public Interest* 23 (1) (2022): 7, 11.

[6] Amy Wax and Philip E. Tetlock, "We Are All Racists at Heart," *The Wall Street Journal*, December 1, 2005, https://www.wsj.com/articles/SB113340432267610972.

[7] See Gregory Mitchell and Philip E. Tetlock, "Antidiscrimination Law and the Perils of Mind Reading," *Ohio State Law Journal* 67 (1) (2006): 1023, 1056, identifying concerns about "internal validity" (causation) and "external validity" (applicability, real-world circumstances). For a reply to this "junk science" critique, see Kang and Lane, "Seeing through Colorblindness," 504–509.

[8] Frederick L. Oswald, Gregory Mitchell, Hart Blanton, et al., "Using the IAT to Predict Ethnic and Racial Discrimination: Small Effect Sizes of Unknown Societal Significance," *Journal of Personality and Social Psychology* 108 (4) (2015): 562, 569.

[9] *Jones v. Nat'l Council of YMCA*, 2013 WL 7046374, *9 (N.D. Ill. 2013) (report of Magistrate Judge Arlander Keys).

[10] *Karlo v. Pittsburgh Glass Works, LLC*, 2015 WL 4232600, *7 (W.D. Penn. 2015), *affirmed*, 849 F.3d 61 (3d Cir. 2017) (affirming on narrower grounds).

[11] Robert P. Abelson, "A Variance Explanation Paradox: When a Little is a Lot," *Psychological Bulletin* 97 (1) (1985): 129, 129–130.

[12] See ibid., 131, reporting variance explained as .00317, the square root of which is equivalent to $r = .056$.

[13] David C. Funder and Daniel J. Ozer, "Evaluating Effect Size in Psychological Research: Sense and Nonsense," *Advances in Methods and Practices in Psychological Science* 2 (2) (2019): 156, 161.

[14] See Robert Rosenthal, "How Are We Doing in Soft Psychology?" *The American Psychologist* 45 (6) (June 1990): 775.

[15] See Anthony G. Greenwald, Mahzarin R. Banaji, and Brian A. Nosek, "Statistically Small Effects of the Implicit Association Test Can Have Societally Large Effects," *Journal of Personality and Social Psychology* 108 (4) (2015): 553, 558.

[16] Ibid., 558.

[17] See Jerry Kang, Mark Bennett, Devon Carbado, et al., "Implicit Bias in the Courtroom," *UCLA Law Review* 59 (5) (2012): 1124, 1151.

[18] For evidence of an implicit stereotype in favor of White men versus East Asian men as litigators and how it influences the evaluation of cross-examinations, see Jerry Kang, Nilanjana Dasgupta, Kumar Yogeeswaran, and Gary Blasi, "Are Ideal Litigators White? Measuring the Myth of Colorblindness," *Journal of Empirical Legal Studies* 7 (4) (2010): 886, 900–906.

[19] See Jerry Kang, "What Judges Can Do About Implicit Bias," *Court Review* 57 (2) (2021): 78, 80–81.

[20] See Gregory J. Meyer, Stephen E. Finn, Lorraine D. Eyde, et al., "Psychological Testing and Psychological Assessment: A Review of Evidence and Issues," *The American Psychologist* 56 (2) (2001): 128, 130 (Table 1).

[21] See Greenwald, Poehlman, Uhlmann, and Banaji, "Understanding and Using the Implicit Association Test," 73 (Table 3), finding that implicit attitude scores predicted behavior in the Black/White domain at an average correlation of $r = .24$, whereas explicit attitude scores had correlations of average $r = .12$. See also Kurdi, Seitchik, Axt, et al., "Relationship Between the Implicit Association Test and Intergroup Behavior," 569–586, finding that implicit biases provide a unique contribution to predicting behavior (ß = .14) and does so more than explicit measures (ß = .11).

[22] The Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. §1607.4(D)(2023).

[23] See, for example, Jerry Kang, "Rethinking Intent and Impact: Some Behavioral Realism about Equal Protection," *Alabama Law Review* 66 (3) (2015): 627–651 (2014 Meador Lecture on Equality); and Jerry Kang,"The Missing Quadrants of Antidiscrimination: Going Beyond the 'Prejudice Polygraph,'" *Journal of Social Issues* 68 (2) (2012): 314–327.

[24] For a fuller account, see Kang and Lane, "Seeing through Colorblindness," 490–492.

[25] This essay does not discuss how judges should try to avoid implicit bias in their own decision-making. For analysis and recommendations, see Kang, "What Judges Can Do about Implicit Bias," 78–91.

[26] See *Griggs v. Duke Power Co.*, 401 U.S. 424, 431 (1971) (introducing disparate impact theory for Title VII employment discrimination). This theory of liability was later ratified by Congress in 1991. Civil Rights Act of 1991, Pub. L. No. 102-166, 105 Stat. 1071, 1074 (codified at 42 U.S.C. § 2000e-2[k]).

[27] See *Washington v. Davis*, 426 U.S. 229 (1976).

[28] Ibid., 248. This interpretation was strengthened three years later in *Personnel Adm'r of Massachusetts v. Feeney*, 442 U.S. 256 (1979), in the context of gender.

[29] I take these gradations from the Model Penal Code § 2.02 (General Requirements of Culpability). *Purposely* means that it is a person's "*conscious* object to engage in conduct of that nature or to cause such a result;" *knowingly* means a person "*is aware* that his conduct is of that nature or that such circumstances exist" or is "*practically certain* that his conduct will cause such a result;" *recklessly* means that a person "*consciously disregards* a substantial and unjustifiable risk" that "involves a gross deviation from the standard of care that a reasonable person would observe in the actor's situation;" *negligently* means a person "*should be aware* of a substantial and unjustifiable risk" that "involves a gross deviation from the standard of care that a reasonable person would observe in the actor's situation." Ibid. [emphasis added]. In the Model Penal Code, "intentionally" or "with intent" means purposely. See MPC § 1.13(12) (Definitions) ("'intentionally' or 'with intent' means purposely").

[30] See Alan David Freeman, "Legitimizing Racial Discrimination Through Antidiscrimination Law: A Critical Review of Supreme Court Doctrine," *Minnesota Law* Review 62 (6) (1978): 1049–1119.

[31] *Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc.*, 576 U.S. 519 (2015).

[32] 42 U.S.C. § 3604(a), § 3605(a).

[33] *Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc.*, (2015), 540 [emphasis added].

[34] *Kimble v. Wisconsin Department of Workforce Development*, 690 F. Supp. 2d 765 (E.D. Wis. 2010).

[35] For early scholarship recommending a category causation standard, see Linda Hamilton Krieger and Susan T. Fiske, "Behavioral Realism in Employment Discrimination Law: Implicit Bias and Disparate Treatment," *California Law Review* 94 (4) (2006): 997, 1053–1054; and Linda Hamilton Krieger, "The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity," *Stanford Law Review* 47 (6) (1995): 1161, 1226.

[36] See 690 F.Supp.2d 765, 768–769 [emphasis added].

[37] Ibid., 775–776 [emphasis added].

[38] Ibid., 776 [emphasis added].

[39] See ibid., 776 (citing articles appearing in the 2006 Behavioral Realism Symposium).

[40] *State of Kansas v. Davon M. Gill*, 56 Kan. App. 2d 1278 (2019).

[41] Kan. Stat. Ann. § 22-4609.

[42] 56 Kan. App. 2d 1278, 1288 (Powell, J., Dissenting).

[43] Ibid., 1286.

[44] Ibid., 1286–1287 [emphasis added].

[45] *Woods v. City of Greensboro*, 855 F.3d 639 (4th Cir. 2017).

[46] Ibid., 641 [emphasis added].

[47] Ibid., 651–652 [emphasis added].

[48] Ibid., 652 [emphasis added].

[49] *Batson v. Kentucky*, 476 U.S. 79 (1986).

[50] Wash. St. Ct. Gen. R. 37; and *Washington v. Jefferson*, 429 P.3d 467 (Wa. 2018).

[51] Ibid., 470. See also Wash. St. Ct. Gen. R. 37(e).

[52] Wash. St. Ct. Gen. R. 37(f) [emphasis added].

[53] Assem. Bill 3070, ch. 318 (Cal. 2020), codified at Cal. Civ. Proc. Code § 231.7, Sec. 1(c) [emphasis added].

[54] Cal. Civ. Proc. Code § 231.7, Sec. 2(d) [emphasis added].

[55] See Order Amending Rules 18.4 and 18.5 of the Rules of Criminal Procedure, and Rule 47(e) of the Rules of Civil Procedure, No. R-21-0020 (Ariz. 2021). Subsequently, there was a legislative effort in Arizona to reinstate peremptory challenges in criminal cases, but this attempt was rejected by Arizona's legislature. See H.B. 2413 (Ariz. 2022).

[56] "Sec 5–12(d)," in *Official 2023 Connecticut Practice Book (Revision of 1998): Containing Rules of Professional Conduct, Code of Judicial Conduct, Rules for the Superior Court, Rules of Appellate Procedure, Appendix of Forms, Notice Regarding Official Judicial Branch Forms, Appendix of Section 1-9B Changes* (Hartford: The Commission on Official Legal Publications, 2023), 180 [emphasis added].

[57] "Sec. 5–12(e)," in *Official 2023 Connecticut Practice Book*, 180 [emphasis added].

[58] N.J. Ct. R. 1:8–3A [emphasis added].

[59] Ibid., Official Comment (3) [emphasis added].

[60] *McCleskey v. Kemp*, 481 U.S. 279 (1987).

[61] Ibid., 312–313.

[62] *Washington v. Gregory*, 427 P.3d 621, 633 (Wash. 2018) (quoting *State v. O'Dell*, 358 P.3d 359 [2015]).

[63] Ibid.

[64] Ibid., 635, [second emphasis added].

[65] Doing the same for Connecticut's death penalty, see also *Connecticut v. Santiago*, 318 Conn. 1 (2015).

[66] *Comcast Corporation v. National Association of African American-Owned Media, et al.*, 140 S.Ct. 1009 (2020), 1014.

[67] For thoughtful analysis, see Katie Eyer, "The But-For Theory of Antidiscrimination Law," *Virginia Law Review* 107 (8) (2021) : 1621.

Retooling Career Systems to Fight Workplace Bias: Evidence from U.S. Corporations

Alexandra Kalev & Frank Dobbin

The civil rights movement spurred U.S. companies and universities to implement antidiscrimination programs. Beginning in the early 1960s, employers adopted antibias training as their first line of defense against bigotry. Even then, there was substantial evidence that this approach was unlikely to lessen bias. In this essay, we discuss social science research on the effects of antibias training, as well as research on systemic approaches to reducing institutional discrimination based on insights from contact theory. As sociologist Samuel Stouffer and psychologist Gordon Allport, the progenitors of contact theory, might have predicted by the end of World War II, we find that interventions to change career systems to maximize intergroup contact can promote workplace equity.

Civil rights protests of the 1950s and 1960s led to new laws against discrimination, and the rapid spread of workplace antibias training programs. When John F. Kennedy directed federal agencies and companies with federal contracts to take "affirmative action" to stop discrimination in 1961, many began with "race relations workshops" to counter bigotry. Western Electric's mass trainings included filmed lectures by James Baldwin, Martin Luther King Jr., and Malcolm X, and live speeches by Roy Wilkins of the NAACP and Whitney Young Jr. of the Urban League. The Department of Health, Education, and Welfare introduced equal opportunity training for its three thousand managers, and at the Social Security Administration, fifty thousand staffers had completed training by the end of 1971. By 1976, more than 60 percent of America's big companies had instituted training programs for managers.[1]

In the 1980s, Ronald Reagan rolled back civil rights regulations and appointed conservative Clarence Thomas to be chair of the Equal Employment Opportunity Commission, a position tasked with enforcing the 1964 Civil Rights Act. Antibias trainers thought the end of affirmative action law was in sight and sketched a business case for inclusion, arguing that women and people of color would soon be the backbone of the workforce and that firms would therefore need to fight discrim-

https://doi.org/10.1162/daed_a_02056

ination to prosper. Consultants heralded antibias training as good management practice. Yet because lawsuits did not abate, corporate counsel still sold antibias training as a means to fend off lawsuits, and plaintiff lawyers still asked for it in discrimination settlements.[2] Attorneys placed faith in it. Trainings have taken different forms over time and, over the last two decades, they have increasingly covered ideas from implicit bias research, both in bespoke live training sessions for corporate leaders and in online trainings for frontline workers. But the core idea behind training has changed little : bias is to blame for workplace inequalities, and bias can best be corrected with self-awareness through training.

While many managers and lawyers see training as a panacea for workplace bias, social scientists have known for decades that efforts to reduce intergroup animus through training typically fail. In 1945, the Social Science Research Council created a Committee on Techniques for Reducing Group Hostility in response to the rise of Hitler, Mussolini, Stalin, and the Klan. In 1947, Cornell sociologist Robin Williams, Jr. surveyed scores of bias reduction efforts for the committee, finding fourteen that used pre/post comparisons or control groups to assess trainings to reduce white people's bias against Black people.[3] Five showed plausible, albeit small, positive effects. But the training with the clearest effect was not one designed to reduce bias, but rather one to *increase* existing bigotry.[4] Williams concluded that it is difficult to extinguish racial bigotry in trainings.

We know a lot more now than we did after Robin Williams's 1947 review. In 2009, psychologist Elizabeth Levy Paluck and political scientist Donald Green published a review of 985 studies of antibias education efforts in schools, universities, nonprofits, and corporations. Where they were subjected to credible tests, training sessions had weak or null effects on bias, and few had lasting effects.[5]

Studies that explore training effects on implicit bias have since proliferated and, in 2019, implicit bias researcher Patrick Forscher and colleagues published a network meta-analysis of 492 such studies with nearly ninety thousand participants. Training can reduce implicit bias in the short term, but effect sizes are small and training does not reduce explicit bias or actual discrimination. Trainings that invoke personal motivations were more likely to reduce implicit bias, while those that invoke threats were less likely.[6] The threat of legal sanction also appears to backfire in real-world diversity training initiatives. In their 2022 review of the literature on diversity trainings in workplaces, psychologists Patricia Devine and Tory Ash similarly found small or no effects.[7] In a recent large-scale field experiment of antibias training, social psychologist Edward Chang and colleagues also found that while for some groups training can decrease measured bias, it does not necessarily reduce discrimination.[8] The research record on antibias training is not promising.

The interventions explored in these studies represent the best case. They were designed by psychologists based on past research about the drivers of attitudinal

change and carried out by scholars attuned to avoiding exposure to situations that might taint results. In real life, diversity trainings are often developed by compliance experts who believe, for instance, that the threat of sanction drives behavior. Those trainings signal in myriad ways that the threat of lawsuit is the reason for training: sessions are run by compliance departments that mandate attendance, highlight the risk of litigation, and test trainees on the law.

Why might such antibias initiatives fail even under the best of circumstances? A number of studies point to possible reasons. First, workplace training rarely changes people if it is not accompanied by changes in work systems and routines – even safety training designed to protect workers. One survey of eighty safety training programs found that only twenty-five changed behavior and only seventeen reduced injuries. Moreover, the features of safety programs that made them effective (trainings that are run live, across multiple sessions, in small groups, and with practice of new routines) are rare in antibias trainings.[9] Second, training to reduce stereotypes tends to activate them. Please do not think of elephants.[10]

Third, training can make trainees think that the problem has been solved. Thus, social psychologist Cheryl Kaiser and colleagues found that when subjects are told that their employers have prodiversity measures, such as training, they presume that the workplace is free of bias and react harshly to claims of discrimination.[11] Similarly, sociologists Emilio Castilla and Stephen Benard found that when people are told their workplaces are free from bias, they become less likely to censor their own biases.[12]

Fourth, social psychologist Victoria Plaut and colleagues found that diversity messaging can make white people feel left out or think they will not be treated fairly, and lead them to oppose equity initiatives.[13] This may be why white workers often leave training feeling "confused, angry, or with *more* animosity toward" other groups.[14] Finally, a large body of research shows that people react negatively to efforts to control their behavior, including efforts to reduce prejudice. As it turns out, white subjects resent external pressure to control prejudice against Black people. When asked to control their own biases, they respond by unleashing them.[15]

While evidence against the efficacy of diversity training might be expected to hearten conservatives, some use it to argue for abolishing diversity, equity, and inclusion (DEI) efforts altogether.[16]

Sociologists have long argued that restructuring work to maximize contact might be more promising than antibias training. The first good evidence that you could eliminate stereotypes and animus between groups by increasing intergroup contact at work comes from the European battlefield in World War II. Harvard sociologist Samuel Stouffer was leading a team commissioned by the U.S. federal government to study soldier adaptation to war when a change in U.S. Army policy set up an unplanned experiment.

The Armed Forces were segregated during the war; Black soldiers and white soldiers did not work together. But as the war progressed, General Dwight D. Eisenhower faced shortages of combat troops in Europe. Eisenhower had to fill in all-white companies after soldiers were lost in battle. He decided to use Black replacement platoons when there were not enough white platoons available. The policy created a natural experiment. Platoons of twenty to forty soldiers were never integrated, but Black and white platoons worked side by side in integrated companies made up of three or four platoons. In some of the white companies that got Black reinforcements, most soldiers were from the Jim Crow South.

Stouffer's team of sociologists, including Shirley Star and Robin Williams, Jr., set out to examine how Eisenhower's solution would affect white soldiers' attitudes toward Black soldiers. They surveyed white soldiers whose companies had been joined by Black platoons and those whose companies had not, posing a question: "Some Army divisions have companies which include Negro platoons and white platoons. How would you feel about it if your outfit was set up something like that?" White soldiers from all-white companies overwhelmingly (62 percent) checked "Would dislike it very much," but only 7 percent of White soldiers from integrated companies felt the same. The flip side was that 60 percent of troops from integrated companies said they "would like it" or would "just as soon have it as any other set-up," while only 11 percent of troops from all-white companies agreed.

White soldiers in integrated companies recognized the change themselves: two-thirds reported that they were initially opposed to the idea of integration. Star, Williams, and Stouffer wrote that this shift gave them "some conception of the revolution in attitudes that took place among these men as a result of enforced contacts."[17] They argued that the success of "this experiment" was tied to the fact that attention was focused on "concrete tasks and goals requiring common effort."[18] These men worked as equals against a common enemy.

In 1954, Stouffer's Harvard colleague Gordon Allport published *The Nature of Prejudice*, suggesting that contact between groups reduces prejudice when the two are of equal status, are cooperating toward a common goal, and have institutional support. Fifty years later, more than five hundred studies in over thirty-five countries had confirmed these ideas.[19] Across a wide range of settings, divided by race, ethnicity, and religion, contact at work was found to reduce prejudice and blur group boundaries.[20] But members of different groups have to be working as equals: slavery didn't do it, and as sociologist Rosabeth Kanter found, men and women working together in gender-segregated career lines don't overcome bias either.[21] But as we argue below, it appears that mentoring relationships between people of different ranks, but within career lines, can counter bias.

Can we change employment systems to increase intergroup contact at work, and thereby reduce bias? Research on contact theory has proven, across many

contexts, that when people from different race and ethnic groups work side by side, *as equals*, racial bigotry declines. The problem is that American firms are often highly integrated when viewed from ten thousand feet, but close up, jobs are highly segregated. In many workplaces, white and Asian-American men dominate tech jobs; women of color dominate DEI departments and customer service; and Black and Brown men prevail in logistics. Sociologists James Baron and William Bielby found this was the case in public and private workplaces alike in the 1980s: despite a growth in gender, race, and ethnic diversity at the workplace, jobs remained stubbornly segregated. Sociologist Corre Robinson and colleagues found much the same pattern two decades later.[22] Thus, we may be able to fight bias in the workplace simply by integrating work groups.

Research on contact theory confirms that individual-level bias is reduced by contact. Research, however, has not explored whether systemic changes that foster intergroup contact promote workforce equity. To explore this, we use data on interventions known to increase intergroup contact, and subsequent data on the diversity of the managerial workforce. The diversity of managers, we believe, is the best measure of workplace equity because management is the hardest, and usually last, level to diversify.

Here we summarize findings from the research literature, including findings from our recent book *Getting to Diversity: What Works and What Doesn't*. There we analyze data from over eight hundred firms, with eight million workers, for the years 1971 to 2015.[23] We assess effects of dozens of different employment systems and programs on the diversity of managers. Our analyses, in essence, compare the share of managers from different race, ethnic, and gender groups in the years before and after each program is in place. We control for diversity of the industry and state labor forces, and among the firm's own nonmanagers, as well as other firm features known to affect diversity (other diversity programs, DEI offices, HR policies and programs).

By following many firms over time, we can isolate the effects of individual programs: if one firm adopts three programs in a single year, there will be plenty of other firms that adopt those very programs at different times. Robustness tests give us confidence that we have identified effects of specific programs and not something else that happens at the same time, such as the arrival of a new CEO committed to DEI. The results are promising.[24]

The advantages our data hold – of permitting us to isolate effects of particular programs on management diversity – highlight the challenges that managers face in assessing their own programs. Managers do not operate in a sterile lab. Many things outside of their control shape diversity: labor supply, recessions, headhunters, and the rise of online job boards, to name a few. Many things happening within the firm shape diversity, including changes in recruitment, promotion,

layoff, and diversity programs. With a huge dataset, we can control for all of these variables. But chief diversity officers with the best of data on their own firms cannot control for these things and, as a result, they rarely have a clue about whether their programs are actually working.

What can these rich data tell us about the effects on managerial diversity of training programs and programs that increase intergroup contact? In our research on diversity training, we test the proposition that the most common form of diversity training – legalistic training for managers – backfires, leading to decreases in the diversity of managers. That is exactly what we have found. Moreover, we have found that curriculum is what divides the worst from the best. Since the 1960s, virtually all programs have covered bias. Beyond that, the worst curricula focus on legal compliance, the best on cultural inclusion.

Legal-compliance training details what the law forbids and how managers can avoid lawsuits, signaling that litigation prevention is goal number one. Often training begins with stories of well-known lawsuits to pique manager interest. An electronics industry HR manager who designed his firm's training told us, "It's always an eye-opener. They love to hear about the latest lawsuits."[25] An Atlanta hospital covers situations that might spark suits; trainers present vignettes and ask managers to discuss how they would handle the situation "in order to keep the organization away from . . . liability." At a Chicago food firm, an HR specialist explains, the main objective of training is to convey "what not to do as a supervisor or manager" and what steps to take if something is reported.

The format of training also signals that lawsuit avoidance is the main goal. Thus, trainings are usually mandatory, offered by compliance or legal departments, and end with exams on the content. A California hospital concludes its annual online manager training with nine exam questions: "In order to pass it you have to get nine out of nine. It won't let you finish if one is wrong." Trainees are told that HR saves their exam results so if there is a complaint naming them, the firm can prove it had done its part explaining the rules to them.

Is this kind of training effective? One laboratory study of MBA students found that white students were more resistant to training when legal compliance was the express motive than when improving performance was the motive.[26] In our analyses, we found significant negative effects of introducing legal-compliance training on the share of five groups in management: white women (12 percent), Black men (5 percent), Black women (14 percent), Asian-American men (7 percent), and Asian-American women (5 percent). For Hispanic men and women, coefficients were negative but not significant.[27] Some of these groups hold few management positions to begin with, so a 5–14 percent decrease does not amount to a big change in the composition of management. But antibias training is not supposed to decrease management diversity. We also found negative or null effects of

legalistic diversity training on the diversity of employees at other levels in these firms.

Legalistic diversity training for managers is designed to make managers aware of their own biases and to make it clear that the law requires them to stop acting on those biases. It follows a simple logic, but the evidence suggests that the underlying idea is wrong. Unfortunately, three-quarters of firms that train managers use a legal-compliance approach.

What about training that does not mention the law? Cultural-inclusion diversity training usually begins with an introduction to implicit bias, but the express goal is to improve communication and collaboration across groups rather than to prevent litigation. Trainers emphasize that good managers know how to handle diversity on their teams and how to foster teamwork. The message is positive: all workplaces are becoming more diverse, and ours will flourish if our managers can create an environment in which everyone can work to their maximum potential.

Cultural-inclusion training dates to the 1960s, but today only about one-quarter of firms offer it without the legalistic content that can poison the well. It usually begins with an invitation from a DEI officer, not a compliance official. One tech firm told us that managers are invited to attend and asked to RSVP, not ordered to attend. The training is often led by a coworker trained as a facilitator, never by a lawyer. "Compliance" isn't in the course description. One health care organization calls its optional inclusion program "Valuing Differences." A Bay Area tech firm touts its program as "an experiential . . . manager training around leading a multicultural team." A large food-processing company emphasizes the importance of engaging everyone on the team: "It's helping managers build their toolkits. . . . They walk out with an engagement action plan – what am I going to do differently in my job? I am going to interact with team members [differently], and I know how to see things differently."

The point of cultural-inclusion training is to teach managers to listen to people from different backgrounds, so as to understand their challenges, and to observe interactions of their workers, so as to understand their experiences. Managers learn how to integrate everyone on the work team so each can work to their potential.

This type of training is akin to harassment training for managers, which commonly works on listening and observational skills to help trainees understand the diversity of worker experiences. Both kinds of training offer lessons in managing people from different groups. We do not have data to distinguish between legalistic and listen-and-observe harassment training. Nonetheless, we find that both cultural-inclusion diversity training for managers and the combined types of harassment training for managers have positive effects.

Virtually all employers who use cultural-inclusion training for managers also run harassment training. In our analyses, cultural-inclusion training picks up the effects on non-white men, and harassment training picks up the effects on all groups of women. Thus, firms that employ cultural-inclusion training for managers see significant subsequent increases in the share of managers who are Black men (10 percent), Hispanic men (14 percent), and Asian-American men (8 percent). White men see a corresponding decline (12 percent). Firms that add sexual harassment training for managers see significant increases in the share of managers who are white women (5 percent), Black women (4 percent), Hispanic women (3 percent), and Asian-American women (3 percent). On average, we observe effects of these trainings for seven years: we are seeing *sustained* positive effects for cultural-inclusion training.

Cultural-inclusion and harassment training for managers can reduce workplace inequality, and so if all firms stopped offering legalistic diversity training for managers and offered these trainings instead, we might see increases in the diversity of managers. But even the best of training will not produce workplace equity anytime soon. As we will see, changes in employment systems designed to foster intergroup contact have the potential to speed up progress considerably.

How effective are programs that foster intergroup contact through changes in work systems? We consider the effects of four programs: targeted recruitment, cross-training, formal mentoring, and self-managed work teams. By creating personal connections between people from different identity groups, these programs activate the surest mechanism social science has identified to quash bias: intergroup contact among coworkers.

Of course, these programs may also increase workplace equity by their intended means: of helping firms to recruit, train, mentor, and manage people from all backgrounds. But what these programs have in common is that they increase contact between groups and are, we have discovered, more effective than most of the many other programs firms have experimented with.

Targeted recruitment programs extend the reach of traditional recruitment systems, which were developed a century ago for the recruitment of white men, to women and people of color. Many firms have changed recruitment strategies little in recent decades. They advertise in generic venues: previously, major newspapers; now, popular job boards. They actively recruit at the alma maters of their managers – often Big Ten and Ivy League schools – neglecting Historically Black Colleges and Universities (HBCUs), Hispanic-Serving Institutions (HSIs), and women's colleges. Many neglect the professional associations of Black, Hispanic, Indigenous, LGBTQ+, and women engineers, lawyers, physicians, and MBAs.

They often fail to see that they are missing important talent pools in doing so. As one Black nurse explained to sociologist Adia Wingfield, the devaluation of

HBCUs makes it hard for Black nurses to find jobs: "I think employers really tend to focus on the school that you went to, your grades, your references . . . in particular, the school because the school makes a huge impact on what the employer sees as a good nurse or not such a good nurse."[28] Historically Black colleges are not on the radar of the many recruiters for nursing jobs who attended historically white colleges.

Our interviews suggest that targeted recruitment helps to promote workforce diversity by creating intergroup contact for both white men managers and the recruits they bring in. At one Boston food industry firm, the HR director reports, the "biggest successes" in recruiting diverse staff came from a decision to send managers to local colleges with large non-white student bodies. "It's usually an open session," she said, where students sit down to talk with three or four managers from different units, one after another. If a manager finds a potential hire, they invite her to visit headquarters. Managers who find recruits themselves precommit to making the hire work, unlike managers whose recruits come through third-party recruitment channels.

The new recruit thus arrives with a manager on their side, typically a manager from a different background or identity group, often white and male. At a Bay Area tech firm, targeted recruitment is done at local Hispanic-serving schools by people from departments that are hiring. Prospects are invited to the facility to meet with half a dozen team members individually. Those hired come with the backing of many on the team they will join.

Thus, targeted recruitment activates the positive effects of intergroup contact in both directions. White men managers get to know new recruits from other groups as individuals even before they start work. In turn, people recruited through these programs get to know a supervisor as a supporter even before they start.

While ever-popular legal-compliance diversity training programs lead to reductions in management diversity, targeted recruitment programs lead to increases. We asked firms if they have *any* targeted recruitment programs for women or people of color, and if they do, we asked when they put those programs in place. On average, after firms began targeted recruiting, representation in management showed statistically significant increases for white women (6 percent), Black men (11 percent), Black women (10 percent), Hispanic women (4 percent), Asian-American men (8 percent), and Asian-American women (7 percent). Moreover, targeted recruitment in firms with employee resource groups have proven even more effective, as employee resource groups take charge of making sure recruitment happens. In those firms, we see large positive effects for Black men (23 percent), Black women (22 percent), Hispanic men (9 percent), Hispanic women (15 percent), Asian-American men (12 percent), and Asian-American women (21 percent).

Targeted recruitment surged in popularity among big firms with federal contracts after President Kennedy ordered them to take affirmative steps to end discrimination in 1961. By the late 1960s, Fortune 750 CEOs overwhelmingly favored special recruitment for Black workers.[29] But these programs waned during successive recessions, when recruitment is typically curtailed. Today, our research suggests, only about 20 percent of medium and large employers target women and people of color in their recruitment programs. If it became the norm, our analyses suggest, corporate diversity in America might grow quickly.

How do we keep up intergroup contact after onboarding new workers? Once hired, many employees encounter highly segregated work environments. One way to increase intergroup contact after hiring is through rotational training programs, in which new workers move through jobs to acquire different skills. In American companies, this sort of "cross-training" is part of a "high-performance workplace" toolkit, designed to maximize skill development and make the workplace more flexible in the face of rapid change, whether in software development, biotech engineering, or steel mini-mills.

While cross-training does not come under the DEI umbrella, it can help break down siloes that, in a typical firm, may cluster women in sales, white men in management, and Black and Hispanic men in production or logistics. In his study of cross-training in manufacturing, sociologist Steve Vallas found that women were eager to rotate into the high-wage jobs usually done by men, gaining new skills that could lead to promotions.[30] It helped them to break out of their siloes. In a clothing factory, sociologist Ian Taplin found that supervisors came to appreciate the abilities of Hispanic employees relegated to low-skill jobs after seeing them master new positions.[31] That's a textbook example of intergroup contact reducing bias.

In our analyses, the introduction of cross-training programs leads to statistically significant increases in the representation of white women (5 percent), Black men (4 percent), Black women (4 percent), Asian-American men (7 percent), and Asian-American women (5 percent) in management. As these groups rise, white men see a corresponding decline (5 percent), as do Hispanic men (4 percent). The negative effect for Hispanic men may be a result of the fact that they are the group least likely to have completed college, and thus are less likely to have jobs that are included in cross-training programs.[32]

How do firms sustain intergroup contact after training? Mentoring programs have the potential to boost intergroup contact by connecting newcomers with higher-ups from different groups. In the absence of formal mentoring, up-and-coming white men more often have mentors than up-and-coming women and people of color, and their connections with their mentors tend to be closer.[33] Thus, when Sun Microsystems created a formal mentoring program open to all, women and people of color signed up in droves. White men more often said, in effect,

"Thanks anyway, I'm good."[34] Firms can both increase the equity of mentoring benefits and promote intergroup contact through formal mentoring programs.

To boost intergroup contact, professional mentoring programs should exhibit three features. First, formalization. Informal mentoring relationships rarely cross gender and racial lines.[35] As management scholar and Morehouse College president David Thomas has shown, when men mentor women informally, the relationship may be misperceived as sexual. Men may conclude that it is best to avoid mentoring women.[36] The #MeToo movement may have helped to check harassment at work, but surveys show that it heightened the discomfort men feel in mentoring women informally, exacerbating the problem of a lack of mentorship for women.[37]

Second, to boost intergroup contact, mentoring programs need to match people based on interests, not demographics. The numbers usually give them no choice. In corporate America there are four junior white managers for every senior white manager and twenty-four junior Black managers for every senior Black manager.[38] There is no reason not to match people from different groups. While protégés report stronger social support from same-race mentors, people of color don't advance faster under same-race mentors.[39] And one study found that women matched with mentors by formal programs were 50 percent more likely to be promoted than women who found mentors on their own.[40] That may be because programs connect protégés to higher-ups outside of their normal circles who can provide new opportunities.[41]

And third, to boost intergroup contact, employers must open mentoring programs to workers at all ranks. Many firms have mentoring programs for the top 1 percent, the "high potentials" nominated by executives. Protégés in those programs already have a sponsor in the executive who nominated them. "HiPo" programs may serve a role, but they need to be accompanied by programs that offer everyone a mentor: mentoring has the biggest payoff for people who are not already on an executive's radar.[42] Sun Microsystems found that protégés who hadn't been tagged as stars improved the most with mentors.[43]

We found that the creation of formal mentoring programs led to a statistically significant increase in the representation of Black women (15 percent), Hispanic men (7 percent), Hispanic women (17 percent), Asian-American men (14 percent), and Asian-American women (17 percent) in management. When we focus on industries in which most people have college degrees – electronics and chemical manufacturing – mentoring boosts white women and Black men managers as well, by more than 20 percent.

Contact theory suggests that interactions between people of equal rank help to fight bias. Our findings on mentoring suggest that contact between people of different ranks in the same career line helps to promote workforce equity, likely in part by undermining bias.

Beyond targeted recruitment, cross-training, and formal mentoring programs, can firms build sustained intergroup contact – across segregated jobs and departments – into their everyday work routines and operations? Many firms now use self-managed teams to plan, coordinate, and carry out work across jobs and units. These teams have the added benefit of forging intergroup contact between white people and men, concentrated in certain departments and roles, and people of color and women, concentrated in other departments and roles.

On self-managed teams, employees in different jobs, who otherwise might work under different supervisors and have little interaction with one another, work together to manage their own tasks, with no formal leader. Each team member has a say in the decision-making and coordination that is usually done by managers.[44] To design new products, one tech firm creates self-managed teams made up of engineers, production line technicians, and administrative assistants who meet several times a week to apportion tasks, try out new ideas, and track progress.[45] In a bank, team members in customer service, tech, and administrative roles share responsibility for technical tasks and phone service, jointly managing scheduling, training, and quality control.[46] In a paper mill, workers from a wide variety of jobs collectively plan activities, assign and rotate tasks among themselves, and take charge of production, quality, and safety.[47] These opportunities can be especially important for Black workers, who are more often isolated in segregated jobs and more often face negative stereotypes about their soft skills.[48]

Like cross-training, self-managed teams spread across U.S. firms as part of the high-performance toolkit. Business scholars have produced a spate of studies demonstrating that self-managed teams outperform traditional, hierarchical management in a range of industries, from steelmaking to tech to retail.[49] In consequence, some four in ten firms use self-managed teams to perform some of their core tasks.[50] When we ran the numbers, we found that firms that introduced self-managed teams for at least some of their core production or service tasks saw statistically significant increases in the share of managers who are white women (6 percent), Black men (3 percent), and Black women (3 percent), and a corresponding decrease in managers who are white men (8 percent). Thus, self-managed teams seem to outperform hierarchical management not only when it comes to productivity, but when it comes to equity and inclusion. They might work even better if, as with mentoring programs, all employees were asked to participate.

Already by 1950, social science research had suggested that antibias training was ineffective. Nonetheless, when federal regulations outlawed workplace discrimination in the early 1960s, many leading firms pursued antibias training as a first line of defense. Laboratory and field research on implicit bias training has taken off in recent decades. Academics have developed training

protocols based in the science, yet meta-analyses of their own scientific studies are clear: this sort of training does not reliably reduce implicit bias in the long term or discrimination in the short term.

In our studies of the effects of real-world diversity training programs on actual workforce diversity, results are mixed. Most trainings involve a smidgeon of antibias content. On top of that, trainings typically cover legal compliance, cultural inclusion, or both. Legal-compliance curricula cause training to backfire, leading to reductions in the diversity of the managerial workforce. Cultural-inclusion curricula actually promote management diversity, so long as they are not tainted by legal-compliance material. That finding is promising because only about one-quarter of trainings for managers take this form, suggesting that mass conversion to cultural-inclusion training could provide a significant boost to workforce equity. But even the best type of training only goes so far. There are more effective ways to promote change.

The earliest research on intergroup contact, which suggests that contact can reduce racial animus, points to a promising means of further promoting workplace equity. Since Stouffer's World War II study of integrated army companies, hundreds of studies conducted in many different contexts, with many different groups, have replicated the finding that bringing people from different groups together to work toward common goals with institutional support can reliably reduce bias.

Our research on workplace programs to promote intergroup contact began with the observation that in workplaces with diverse workforces, work groups are seldom integrated. We explored the effects of targeted recruitment of women and non-white workers, cross-training for employees, formal mentoring programs, and self-managed work teams. Of these four management innovations, only targeted recruitment is a diversity program. The others are popular for improving management generally. All four programs, however, significantly increase intergroup contact. Targeted recruitment sends managers, including white men, to recruit people from colleges and professional associations serving women and people of color. Those managers become mentors and sponsors of the people they bring on board. Cross-training rotates employees through different departments for a month or two each, giving new recruits working in departments that are largely segregated by sex and race sustained contact with people from other groups. Formal mentoring programs typically extend mentoring to women and people of color, creating new contacts with white male managers. Self-managed teams bring together people from different roles, and at different levels, as equal team members to manage production or service provision without a leader.

These approaches work remarkably well at promoting workplace equity, as measured by the diversity of managers. Changes in employment systems, together, can do significantly more than even the most effective diversity and harassment training programs to promote workplace equity. With the data now avail-

able, we cannot know whether these effects result principally from reductions in employee bias. But we do know from hundreds of previous studies that intergroup contact at work reduces bias, so it stands to reason that these changes in employment systems, which are known to promote intergroup contact, reduce bias and promote workforce diversity.

The finding that work systems that increase intergroup contact increase equity also provides insight into hierarchical organizations more broadly. While hierarchical management may be efficient in certain contexts, it tends to reproduce status inequalities and strengthen out-group biases.[51] The rich evidence about the importance of collaborative contact from studies testing contact theory, and from our own research on corporate DEI programs, suggests that workplaces with rigid hierarchies, such as universities and law firms, may face challenges in reducing implicit biases to promote equity.

We have noted that the most popular form of diversity training – legal-compliance training – often leads to backlash and reductions in the diversity of managers. Here, we have focused on the promise of systemic changes to promote equity. Does antibias training have a role to play? Our research suggests that it does, but by itself, current legalistic forms of antibias training are unlikely to promote equity. On its own, antibias training that teaches listening and management skills for cultural inclusion does promote equity.[52] Moreover, we find that even legalistic diversity training can augment the positive effects of systemic changes. When such training is introduced with mentoring or employee resource groups, it renders them more effective. And when such training is introduced in tandem with measures to apportion responsibility for equity and inclusion, such as diversity taskforces or managers, it can boost their effects.[53] Thus, there is a role for antibias training in efforts to promote equity, but we caution that by itself, even cultural-inclusion diversity training is unlikely to move the needle by much.

ABOUT THE AUTHORS

Alexandra Kalev is Associate Professor and Chair of the Department of Sociology and Anthropology at Tel-Aviv University. She is the author of *Getting to Diversity: What Works and What Doesn't* (with Frank Dobbin, 2022) and has recently published in *Work, Employment & Society* and *Harvard Business Review*.

Frank Dobbin is the Henry Ford II Professor of the Social Sciences and Chair of the Sociology Department at Harvard University. He is the author of *Getting to Diversity: What Works and What Doesn't* (with Alexandra Kalev, 2022), *Inventing Equal Opportunity* (2009), and *Forging Industrial Policy: United States, Britain, and France in the Railway Age* (1994).

ENDNOTES

[1] Frank Dobbin, *Inventing Equal Opportunity* (Princeton, N.J.: Princeton University Press, 2009).

[2] Erin Kelly and Frank Dobbin, "How Affirmative Action Became Diversity Management: Employer Response to Antidiscrimination Law, 1961–1996," *American Behavioral Scientist* 41 (7) (1998): 960–984.

[3] Robin M. Williams, Jr., *The Reduction of Intergroup Tensions: A Survey of Research on Problems of Ethnic, Racial, and Religious Group Relations* (New York: Social Science Research Council, 1947), vii.

[4] Ibid., 29.

[5] Elizabeth Levy Paluck and Donald P. Green, "Prejudice Reduction: What Works? A Review and Assessment of Research and Practice," *Annual Review of Psychology* 60 (1) (2009): 339–367.

[6] Patrick S. Forscher, Calvin K. Lai, Jordan R. Axt, et al., "A Meta-Analysis of Procedures to Change Implicit Measures," *Journal of Personality and Social Psychology* 117 (3) (2019): 522–559.

[7] Patricia G. Devine and Tory L. Ash, "Diversity Training Goals, Limitations, and Promise: A Review of the Multidisciplinary Literature," *Annual Review of Psychology* 73 (2022): 403–429.

[8] Edward H. Chang, Katherine L. Milkman, Dena M. Gromet, et al., "The Mixed Effects of Online Diversity Training," *Proceedings of the National Academy of Sciences* 116 (16) (2019): 7778–7783.

[9] Michael J. Colligan and Alexander Cohen, "The Role of Training in Promoting Workplace Safety and Health," in *The Psychology of Workplace Safety*, ed. Julian Barley and Michael R. Frone (Washington, D.C.: American Psychological Association, 2004), 223–248.

[10] Mary Lou Egan and Marc Bendick, Jr., "Combining Multicultural Management and Diversity into One Course on Cultural Competence," *Academy of Management Learning & Education* 7 (3) (2008); Adam D. Galinsky and Gordon B. Moskowitz, "Perspective Taking: Decreasing Stereotype Expression, Stereotype Accessibility, and In-Group Favoritism," *Journal of Personality and Social Psychology* 78 (4) (2000); Neil C. Macrae, Galen V. Bodenhausen, Alan B. Milne, et al., "Out of Mind but Back in Sight: Stereotypes on the Rebound," *Journal of Personality and Social Psychology* 67 (5) (1994): 808–817; and Carol T. Kulik, Elissa L. Perry, and Anne C. Bourhis, "Ironic Evaluation Processes: Effects of Thought Suppression on Evaluations of Older Job Applicants," *Journal of Organizational Behavior* 21 (6) (2000): 689–711.

[11] Cheryl R. Kaiser, Brenda Major, Ines Jurcevic, et al., "Presumed Fair: Ironic Effects of Organizational Diversity Structures," *Journal of Personality and Social Psychology* 104 (3) (2013): 504–519.

[12] Emilio J. Castilla and Stephen Benard, "The Paradox of Meritocracy in Organizations," *Administrative Science Quarterly* 55 (4) (2010). See also Laura M. Brady, Cheryl R. Kaiser, Brenda Major, et al., "It's Fair for Us: Diversity Structures Cause Women to Legitimize Discrimination," *Journal of Experimental Social Psychology* 57 (1) (2015).

[13] Victoria C. Plaut, Flannery G. Garnett, Laura E. Buffardi, et al., "'What About Me?' Perceptions of Exclusion and Whites' Reactions to Multiculturalism," *Journal of Personality and Social Psychology* 101 (2) (2011); and Tessa L. Dover, Brenda Major, and Cheryl R.

Kaiser, "Members of High-Status Groups Are Threatened by Pro-Diversity Organizational Messages," *Journal of Experimental Social Psychology* 62 (2016).

[14] Carol T. Kulik, Molly B. Pepper, Loriann Roberson, et al., "The Rich Get Richer: Predicting Participation in Voluntary Diversity Training," *Journal of Organizational Behavior* 28 (2007); and Rohini Anand and Mary-Frances Winters, "A Retrospective View of Corporate Diversity Training from 1964 to the Present," *Academy of Management Learning & Education* 7 (2008).

[15] Patricia G. Devine, E. Ashby Plant, David M. Amodio, et al., "The Regulation of Explicit and Implicit Race Bias: The Role of Motivations to Respond without Prejudice," *Journal of Personality and Social Psychology* 82 (2002): 835–848. See also Lisa Legault, Jennifer N. Gutsell, and Michael Inzlicht, "Ironic Effects of Antiprejudice Messages: How Motivational Interventions Can Reduce (But also Increase) Prejudice," *Psychological Science* 22 (12) (2011); and Deborah L. Kidder, Melenie J. Lankau, Donna Chrobot-Mason, et al., "Backlash toward Diversity Initiatives: Examining the Impact of Diversity Program Justification, Personal and Group Outcomes," *International Journal of Conflict Management* 15 (1) (2004).

[16] Christopher F. Rufo, "D.E.I. Programs are Getting in the Way of Liberal Education," *The New York Times*, July 27, 2023, https://www.nytimes.com/2023/07/27/opinion/christopher-rufo-diversity-desantis-florida-university.html; and Jonathan Butcher, "DEI Doesn't Work–Taxpayers Shouldn't Pay for It," The Heritage Foundation, January 30, 2023, https://www.heritage.org/education/commentary/dei-doesnt-work-taxpayers-shouldnt-pay-it.

[17] Samuel Stouffer, Edward Suchman, Leland DeVinney, et al., *The American Soldier: Adjustment During Army Life*, vol. 1 (Princeton, N.J.: Princeton University Press, 1949), 595.

[18] Ibid.

[19] Thomas F. Pettigrew and Linda R. Tropp, "A Meta-Analytic Test of Intergroup Contact Theory," *Journal of Personality and Social Psychology* 90 (5) (2006): 751–783.

[20] Herbert Blumer, "Race Prejudice as a Sense of Group Position," *The Pacific Sociological Review* 1 (1) (1958): 3–7; Michael Hechter, "Group Formation and the Cultural Division of Labor," *American Journal of Sociology* 84 (2) (1978): 293–318; and Dora C. Lau, and J. Keith Murnighan, "Demographic Diversity and Faultlines: The Compositional Dynamics of Organizational Groups," *The Academy of Management Review* 23 (2) (1998): 325–340.

[21] Rosabeth Moss Kanter, *Men and Women of the Corporation* (New York: Basic Books, 1977).

[22] James N. Baron and William T. Bielby, "Organizational Barriers to Gender Equality: Sex Segregation of Jobs and Opportunities," in *Gender and the Life Course*, ed. Alice S. Rossi (New York: Aldine de Gruyter, 1985), 233–251; and Corre L. Robinson, Tiffany Taylor, Donald Tomaskovic-Devey, et al., "Studying Race or Ethnic and Sex Segregation at the Establishment Level," *Work and Occupations* 32 (1) (2005): 5–38.

[23] Frank Dobbin and Alexandra Kalev, *Getting to Diversity: What Works and What Doesn't* (Cambridge, Mass.: Harvard University Press, 2022).

[24] Full details on methods, and the models themselves, are available at Frank Dobbin, "Online Methodological Appendix," August 17, 2022, https://scholar.harvard.edu/dobbin/home/HUP2022Supplement.

[25] All quotes, except those with endnote references to other sources, come from our team's interviews with line managers, HR managers, and diversity managers at over one hun-

dred firms in the Boston, Chicago, Atlanta, and San Francisco standard metropolitan statistical areas.

[26] Deborah L. Kidder, Melenie J. Lankau, Donna Chrobot-Mason, et al., "Backlash toward Diversity Initiatives: Examining the Impact of Diversity Program Justification, Personal and Group Outcomes," *International Journal of Conflict Management* 15 (1) (2004): 77–104.

[27] In exploratory analyses, we found that any legal-compliance curriculum in training sessions turns trainees off, leading to adverse effects. So, in these analyses, we include all trainings that contain any compliance components, even when they also cover cultural inclusion.

[28] Adia Harvey Wingfield and Koji Chavez, "Getting In, Getting Hired, Getting Sideways Looks: Organizational Hierarchy and Perceptions of Workplace Racial Discrimination," *American Sociological Review* 85 (1) (2020): 31–57.

[29] Joseph R. Goeke and Caroline S. Weymar, "Barriers to Hiring Blacks," *Harvard Business Review* 47 (5) (1969): 144–152.

[30] Steven P. Vallas, "Empowerment Redux: Structure, Agency, and the Remaking of Managerial Authority," *American Journal of Sociology* 111 (6) (2006): 1677–1717; and Steven P. Vallas, "Why Teamwork Fails: Obstacles for Workplace Change in Four Manufacturing Plants," *American Sociological Review* 68 (2) (2003): 223–250.

[31] Ian Taplin, "Flexible Production, Rigid Jobs: Lessons from the Clothing Industry," *Work and Occupations* 22 (4) (1995): 412–438.

[32] Donald Tomaskovic-Devey, Anthony Rainey, Jasmine Kerrissey, and Steve Boutcher, "Separate and Unequal: The Impact of Segregation and Within Job Disparities on Public Sector Intersectional Earnings Gaps," working paper, 33.

[33] Gail M. McGuire, "Gender, Race, and the Shadow Structure: A Study of Informal Networks and Inequality in a Work Organization," *Gender and Society* 16 (3) (2002): 303–322.

[34] Capital Analytics, *Sun Microsystems University Mentoring Program* (Durham, N.C.: Capital Analytics, 2006).

[35] David A. Thomas, "Mentoring and Irrationality: The Role of Racial Taboos," *Human Resource Management* 28 (2) (1989): 279–290.

[36] Raymond A. Noe, "Women and Mentoring: A Review and Research Agenda," *Academy of Management Review* 13 (1) (1988): 65–78.

[37] LeanIn.org, "Working Relationships in the #MeToo Era," polls conducted in 2018–2019, https://leanin.org/sexual-harassment-backlash-survey-results.

[38] U.S. Equal Employment Opportunity Commission, "Job Patterns for Minorities and Women in Private Industry," data extracted July 6, 2021, https://www.eeoc.gov/statistics/employment/jobpatterns/eeo1.

[39] David A. Thomas, "The Truth about Mentoring Minorities: Race Matters," *Harvard Business Review* 79 (4) (2001): 99–107.

[40] Herminia Ibarra, Nancy M. Carter, and Christine Silva, "Why Men Still Get More Promotions Than Women," *Harvard Business Review* 88 (9) (2010): 85.

[41] Tammy D. Allen, Lillian T. Eby, and Elizabeth Lentz, "Mentorship Behaviors and Mentorship Quality Associated with Formal Mentoring Programs: Closing the Gap between Research and Practice," *Journal of Applied Psychology* 91 (3) (2006): 567–578.

[42] Ellen Berrey, "Breaking Glass Ceilings, Ignoring Dirty Floors," *American Behavioral Scientist* 58 (2) (2014): 347–370.

[43] Wharton School, *Wharton on Learning Leadership, Volume 1: Creating a Learning Environment* (Philadelphia: Wharton School of the University of Pennsylvania, 2007), 6–7.

[44] Peter Cappelli, Laurie Bassi, Harry Katz, et al., *Change at Work* (New York: Oxford University Press, 1997); Paul Osterman, "Work Reorganization in an Era of Restructuring: Trends in Diffusion and Effects on Employee Welfare," *Industrial & Labor Relations Review* 53 (2) (2000): 179–196; and Eileen Appelbaum and Peter Berg, "High-Performance Work Systems and Labor Market Structures," in *Sourcebook of Labor Markets: Evolving Structures of Processes*, ed. Ivar Berg and Arne L. Kalleberg (New York: Kluwer Academic/Plenum, 2001), 271–293.

[45] Jerry Daday and Beverly Burris, "The Effects of Teaming-Structures on Race, Ethnicity, and Gender Differences in a High-Tech Corporation: A Case Study," 12, paper presented at the American Sociological Association Conference, August 16–19, 2002.

[46] Marjukka Ollilainen and Joyce Rothschild, "Can Self-Managing Teams Be Truly Cross-Functional? Gender Barriers to a 'New' Division of Labor," *Research in the Sociology of Work* 10 (2001): 141–164.

[47] Vallas, "Why Teamwork Fails."

[48] Adia Harvey Wingfield and Renée Skeete Alston, "Maintaining Hierarchies in Predominantly White Organizations," *American Behavioral Scientist* 58 (2) (2014): 274–287; and Philip Moss and Chris Tilly, "'Soft' Skills and Race: An Investigation of Black Men's Employment Problems," *Work and Occupations* 23 (3) (1996): 252–276.

[49] Eileen Appelbaum, Thomas Bailey, and Peter Berg, *The New American Workplace: Transforming Work Systems in the United States* (Ithaca, N.Y.: Cornell University Press, 2000); Paul Osterman, "Work Reorganization in an Era of Restructuring: Trends in Diffusion and Effects on Employee Welfare," *Industrial and Labor Relations Review* 53 (2) (2000): 179–196; and Michael J. Handel and David I. Levine, "Editors' Introduction: The Effects of New Work Practices on Workers," *Industrial Relations* 43 (1) (2004): 1–43.

[50] Alexandra Kalev, "Cracking the Glass Cages? Restructuring and Ascriptive Inequality at Work," *American Journal of Sociology* 114 (6) (2009): 1591–1643; and Arne L. Kalleberg, Peter V. Marsden, Jeremy Reynolds, and David Knoke, "Beyond Profit? Sectoral Differences in High-Performance Work Practices," *Work and Occupations* 33 (3) (2006): 271–330.

[51] Joyce Rothschild-Whitt, "The Collectivist Organization: An Alternative to Rational-Bureaucratic Models," *American Sociological Review* 44 (4) (1979): 509–527; and Elizabeth H. Gorman and Sarah Mosseri, "How Organizational Characteristics Shape Gender Difference and Inequality at Work," *Sociology Compass* 13 (3) (2023): e12660, https://doi.org/10.1111/soc4.12660.

[52] Dobbin and Kalev, *Getting to Diversity: What Works and What Doesn't*.

[53] Alexandra Kalev, Frank Dobbin, and Erin Kelly, "Best Practices or Best Guesses? Diversity Management and the Remediation of Inequality," *American Sociological Review* 71 (4) (2006): 589–617.

Implicit Bias versus Intentional Belief: When Morally Elevated Leadership Drives Transformational Change

Wanda A. Sigur & Nicholas M. Donofrio

The twenty-first century is witnessing rapid and deep change in the global economy. These changes require innovation-driven solutions and motivated, skilled workforces. The talents of every person will be required to support performance in every domain, and deliberate actions must be taken to address impediments to full engagement. Even with clear government policy and significant investments in encouraging representation and inclusion of diversity of race, sexual orientation, gender identity, and ability, progress continues to lag. This essay captures promising practices and recommendations for structural or systemic change punctuated with stories of leadership driven by the belief that implementing strategies to disrupt the effects of implicit bias are important to develop diverse, fully engaged populations.

Are there opportunities to shape our future based on the beliefs we articulate? Or are our actions controlled by our unconscious biases, preprogrammed and potentially toxic? For more than fifty years, American polling on perspectives of diverse populations has reflected shifting attitudes. For example, the 2022 Gallup poll on race relations shows a reversal of position on interracial marriage, from 4 percent approval in 1958 to 94 percent in 2021.[1] The poll on values and morality in America indicates a greater acceptance of same sex marriage, from 27 percent to 71 percent approval (1996 to 2022).[2] Psychologists Tessa E. S. Charlesworth and Mahzarin R. Banaji report explicit attitudes on race, sexual orientation, weight, and ability have decreased in bias by 98 percent, 65 percent, 31 percent, and 37 percent respectively since 2007.[3] These supportive trends have been reflected in public commitments to accessibility, inclusion, equity, and diversity by various groups, including the government, industry, and education communities, and reflected in public policy and the media. So why is there still significant evidence of toxicity?

- There are glass ceilings for women, racially minoritized people, and gender-nonconforming people, making them unlikely to hold executive roles.[4]

https://doi.org/10.1162/daed_a_02057

- In companies with more than one hundred employees, Black people make up approximately 3 percent of senior leaders.[5]
- Hiring and promotion for cultural fit is the norm, but only people under forty-five years old are seen as a "good cultural fit" by 85 percent of hiring managers.[6]
- Extroverts and confident talkers are seen as a better cultural fit and more promotable but may not be the best leaders.[7]
- Overweight people are seen as less suitable.[8]
- Explicit evidence of heightened anti-Asian racism during and after the COVID-19 crisis exists.[9]

Apparently, against a backdrop of articulated support, evidence of equality seems limited. Charlesworth and Banaji contrasted explicit attitudes with results from implicit bias testing.[10] Reductions in implicit biases were generally lower than articulated values, sometimes significantly so. Kirsten N. Morehouse and Banaji's essay in this volume provides a detailed discussion.[11]

Let's clarify terminology. *Implicit bias* is a negative attitude, often unconscious, against a specific social group.[12] Other essays in this volume discuss the history, theory, research, and sustainability of interventions of implicit bias.[13] As noted by psychologists Calvin K. Lai and Banaji in their essay on implicit intergroup bias, and Jack Glaser in this volume, unconscious biases may unwittingly produce behaviors that limit access to resources like health care, education, and funding, and may influence workplace decisions.[14] *Intentional belief* refers to moral reasoning, an intentional and conscious mental activity that consists of transforming given information about people (and situations) in order to reach a moral judgment.[15] Initiated by leaders with the intentional belief that they can influence implicit bias by using their "bully pulpit" within their organization to reduce systemic bias, this essay discusses possibilities for sustainable organizational change.

As discussed elsewhere in this volume, individual bias and systemic discrimination are interrelated. Using the "bias of crowds" theoretical framework, Manuel J. Galvan and B. Keith Payne's review of the structural or systemic aspects (used interchangeably) of bias, specifically racism, explains why articulated support of racial equity has not had greater impact. As we address structural inequalities, we reframe the experiences that shape implicit bias.[16] As culture drives experiences, experiences shape bias, bias in society shapes the mind, and, in turn, bias in the mind shapes culture. Institutional and structural bias drives long-standing conditions that reinforce implicit bias. *Both* individual and institutional solutions are needed.[17]

Although significant research exists on implicit bias, evidence of successful interventions is limited. The meta-analyses by psychologists Patrick S. Forscher, Jordan R. Axt, Lai, and colleagues, as well as those by psychol-

ogists Chloe FitzGerald, Angela Martin, Delphine Berner, and Samia Hurst reviewed hundreds of bias-intervention studies and data on tens of thousands of participants.[18] None resulted in long-term change.[19] However, Lai and Banaji acknowledge that the influence of bias may be reduced through various learning processes and practices they recommend. But they also acknowledge that it is "unreasonably optimistic" to assume individuals will choose to change their values on their own on a large scale. Structural solutions may be appropriate at the macro level, with institutions implementing policy and governance solutions that control resources like housing, education, business, and health care, to encourage individuals to close gaps and push for justice and equality.[20] Cass R. Sunstein and Richard H. Thaler suggest a form of "paternal libertarianism" in circumstances such as this, by which they mean that the best approach to organizational decision-making is one within a framework that includes data analysis and a clear cost-benefit analysis.[21]

This essay focuses on systemic solutions. Recognizing the role that work plays in the lives of individuals, and that workplace culture is often a reflection of society, our examples and discussion are focused on leaders and changes in the organization of workplaces. Highlighting promising practices and featuring stories of leaders who have moved beyond their own unconscious biases in favor of actions that elevate their communities, we discuss transformational leaders addressing the effects of implicit bias within their organizations.

Transformational leadership, a theoretical model attributed to historian James MacGregor Burns, highlights that "the result of transforming leadership is a relationship of mutual stimulation and elevation that converts followers into leaders and may convert leaders into moral agents."[22] Building on Burns's work, which focuses on political leaders, psychologist Bernard Bass extended the model to organizational management, adding that transformational leaders "attempt and succeed in raising colleagues, subordinates, followers, clients, or constituencies to a greater awareness about the issues of consequence."[23]

The drivers of transformational leadership originate in personal beliefs and value systems that include such values as justice and integrity. Burns refers to these values as "end values," those that cannot be negotiated or exchanged between individuals. When leaders introduce change at key intervals, it supplies a nudge to redirect and prevent bias-based outcomes.[24]

This model for leadership was selected because of its potential to create and energize the next level of leaders to carry on sustained change, and because of the opportunity to engage individuals, potentially a disagreeing majority, in transforming beliefs. The four dimensions of transformational leadership – inspirational motivation, idealized power, intellectual stimulation, and consideration of individuals – temper the potential for elitist leader-driven change.[25] Sustainable change requires participatory, somewhat democratic methods of engagement, including transpar-

ent communications and commitments from organizational stakeholders, while achieving performance goals. [26] Alexandra Kalev and Frank Dobbin give historical context and multiple examples of organizational change that resulted in a "revolution in attitudes." When common goals are established and members of different groups are working as equals, real change is possible across areas of race, ethnicity, and religion.[27]

Although the pace of change in addressing social issues is inadequate, we know change is possible, particularly when leaders see it as critical and act. Research on transformational leaders who focus on both moral values and data-driven reasoning to drive culture change is limited. However, there are examples of leaders using values to drive decisions, engagement, and inclusion:

- Who hired the first Black woman into a predominantly white male business world and why?
- Who first gave same-sex partners benefits in an assumed heterosexual business world and why?
- Who enables, in a business world rich with choice, all people, no matter who they are, to be their best and do their best each and every day and why?

We explore promising practices for disrupting the effects of implicit bias through reflections on each of our own experiences working for strong leaders with strong beliefs and clear value systems (even if their words may have revealed internal biases to the contrary).

Donofrio: IBM, a company I grew up in and lived in for forty-four years, not only taught me these lessons but actually had a history of teaching others these lessons—if you are only willing to look, listen, and learn.

Starting with Thomas J. Watson Sr., who was basically the founder of the IBM Company as it is known today, I often wondered how he knew what to do when everything around him was changing, if not telling him to do something else. How did he know in the 1920s and 1930s that women had a place in business when everything around him told him otherwise? How did he know that equal pay for equal work—no matter your gender, ethnicity, religion, nationality, or skin color—was the right thing to do? While he wasn't always perfect and right and first, somehow he seemed to have implicit belief versus bias! Net, he led from his beliefs, which perhaps became his bias. Or was it the other way around?

Soon after Watson arrived at IBM in 1914, he developed and released his "Basic Beliefs" for all employees. IBM's three Basic Beliefs, the foundation for the values and culture that guided the company decade after decade, were respect for the individual, superlative customer service, and the pursuit of excellence in all tasks.

Even as I joined IBM in 1964, these Basic Beliefs were on the tip of everyone's tongue and guided everything every IBMer did. In three simple thoughts, Watson told

everyone what IBM would do for them and what IBM expected from them. Beliefs, values, culture. There was very little room for bias.

Clearly, Watson made this happen because he saw a critical void and stepped up to fill it based on his own experiences and beliefs. After all, it was his company, and he was charged with its leadership for the betterment of all stakeholders, employees, clients/customers, partners, investors, and communities.

History would suggest that this top-down move worked wonders for the struggling CTR Company Watson joined in 1914, and the incredible IBM success he left in 1956.

The willingness and determination of leaders to act on values for the good of their teams and the good of their enterprise are critical. Both organizational and individual goals are important. Transformational leadership theory pushes leaders to pursue teamwork, communal respect, and cooperation.[28]

Our concentration on the role of leadership in setting up policy, providing incentives like awards, recognition, and promotions, and enforcing accountability with metrics is based on the links between organizational behavior and organizational culture. Employee engagement is a product of the organizational culture with the "daily experience" crafted by colleagues, peers, supervisors, stakeholders, and the organization.

Culture is evident in the vision, values, and frameworks for the behaviors that prevail. Psychologist Edgar Schein has captured these elements when he defines organizational culture as "a pattern of shared basic assumptions that was learned by a group as it solved its problems of external adaptation and internal integration, that has worked well enough to be considered valid and, therefore, to be taught to new members as the correct way to perceive, think, and feel in relation to those problems."[29] Shared assumptions evolve as the group works to adapt to external conditions and achieve internal integration. Basic assumptions and practices are being taught to new members as the correct way to perceive, think, and feel.[30] The issue of bias is associated with the character of the organization, the character of its leadership, and the conviction with which all are willing to shape the organization on the side of right. Much of the leadership challenge is to hand decision-making back to our rational belief system and wrestle it away from implicit bias.

Schein also finds a unique association with leadership in creating culture. The values of the leader are imposed on the group through process and requirements. With success, the culture develops, including assumptions that are taken for granted, like in-group/out-group dynamics, and other elements of culture. The leader has the responsibility to step outside the culture when assumptions are no longer valid and change is needed. "This ability to perceive the limitations of one's own culture and to evolve the culture adaptively is the essence and ultimate challenge of leadership. Leaders begin the culture creation process and, as we will see, must also manage, and sometimes change culture."[31] Sustainable change re-

quires clear goals, change agents, engagement, metrics, reward structures, resources, and leadership support.

Donofrio: Another approach to ensuring you are connected and relevant as you look to innovate in this multigenerational workforce world is to engage everyone and anyone you can in addressing the problem as you seek to unlock hidden value. In 2002 and 2003, as Sam Palmisano took the reins at IBM as CEO and chair, we focused on utilizing technology to enable an open, collaborative, multidisciplined, global platform focused on the problem in a way that everyone could contribute. We called them "Jams." The Values Jam was Sam's brainchild (with complete support of all of us on his senior leadership team) as he determined that IBM needed to update, refresh, and reassess its original Basic Beliefs as laid out by its founder.

Rather than "handing down" his version, Sam chose to ask IBM's employees to engage and help him determine what those beliefs should be that we would all follow and use to build our twenty-first-century culture. Sam engaged the entire IBM global workforce, over three hundred thousand members strong in one hundred seventy countries. Over a few days, more than half of all IBMers engaged to provide their thoughts and ideas with lots of discussion and debate. Shortly thereafter, when all the results were tallied, Sam simply announced what everyone working together had determined: our chosen new IBM Values are, "Dedication to every client's success. Innovation that matters for the company and the world. Trust and personal responsibility in all relationships."

Much has been written about this massively parallel approach to collaboration and problem-solving (in 2004, scores of colleges and universities wrote about the Jams approach and their results). If values and beliefs set the base for culture, which in turn provides everyone the guidelines for action or inaction, what better way to get everyone, including the leaders, on the same page than through an experience like Values Jam? When everyone is looking and everyone is engaged, is there really any room for bias?

Footnote…Just before I graduated from IBM, we held an Innovation Jam that again utilized the entire IBM workforce, with the addition of their family members and IBM's clients, to help us determine where to invest and how to quicken our path to successful innovation. As Sam led the Values Jam, I led the Innovation Jam. What an innovative way to innovate! Hiding biases when nearly two hundred thousand people are engaged and questioning every move, while not impossible, is highly unlikely.

Customers buying goods, investors investing, employees choosing where to work and how to engage – all are tied to a cycle of innovation, performance, and customer attraction. The challenge of creating value or producing something of value for a customer or the business itself involves balancing the needs and interests of external and internal stakeholders. While businesses have always struggled to increase revenue, differentiate themselves, manage costs,

and delight customers, the pace of innovation has been accelerated to accommodate disruptive business models and a growing global marketplace. The key driver of innovation – the people within the enterprise – must be carefully considered and their engagement painstakingly prioritized. The National Commission on Innovation and Competitiveness Frontiers clearly laid the foundation for all to follow since 2003: "To compete in the next economy requires playing a new innovation game, one whose goal is to boost U.S. innovation tenfold," demanding bold leadership, a global perspective, a whole-of-nation strategy, and appropriate support.[32] To meet this demand, the American workforce should be engaged, skilled, and enabled to contribute.[33] Needless to say, this includes every employee in every role.

Sociologist Bas Hofstra and colleagues' research points to higher innovation rates among demographically underrepresented doctoral candidates.[34] The innovation enterprise is at its best when it fully uses the broadest range of human talent.[35] In their meta-analysis of multiple studies, health care researchers Luis Emilio Gomez and Patrick Bernet found that diverse teams can support better decision-making by introducing difficult, unexpected questions that require resolution.[36]

Sigur: The Return-to-Flight effort following the tragic NASA Space Shuttle Columbia accident included both significant technical and personal challenges. A comprehensive investigation found that catastrophic damage to the shuttle orbiter was caused by the loss of large pieces of shuttle external tank foam insulation during launch. The tank production plant, located in New Orleans, was hit by Hurricane Katrina during the safe hardware redesign effort. Ninety-eight percent of the workforce were affected, most left homeless. The factory was surrounded by water, and a concrete roof had collapsed onto near-complete hardware.

I was asked to lead the recovery effort for Lockheed Martin Space, the external tank contractor. Recovery meant every person was needed, both qualified and qualifiable, with no room for racism, ageism, sexism, or sexualism as we worked together on solutions across every front: finding our team members and families and helping reestablish lives while developing solutions for safe flight against a backdrop of schedule pressure to assist with building the Space Station as geopolitical support waned.

Prompted by NASA Human Factors expert Cynthia Null, I redesigned the makeup of solution teams, giving production-floor practitioners a voice along with engineering, a first. They supported everything as a group: design, build, test at work; and house "gutting," rebuilding, and personal support at home. I realized the unique nature of this effort when I observed a "production huddle" before performing a redesigned tank spray: the practitioners, a nearly all-Black group, and the scientists and engineers, a nearly all-white group—expert peers, discussing a solution that had escaped solving for years. Intergroup relationships flourished and continue to this day, including those across racial lines, as they had helped each other succeed both at work and at reknitting their personal lives. And tank vehicle performance was nearly perfect.

Although it may be difficult to revise individual bias, it is definitely possible to interrupt bad behaviors. Promising practices include identifying and owning clearly defined expectations and practices for a) the daily experiences of employees; b) organizational demographics (hiring and promotions); and c) self-assessment tools to measure progress – everything must be linked to the organization's strategy and goals.[37]

You cannot simply "have" diversity. Improved performance is *possible* with diverse teams, but different perspectives, knowledge, and backgrounds will only lead to the promised breakthroughs when issues of communication and integration can be resolved.[38] Dialogue around expected outcomes and behaviors can have significant impact. Left unresolved, employee turnover prompted by bias and unfair treatment (unfair employee assessments, limited access to key assignments and promotions) has cost U.S. employers $172.4 billion over the past five years.[39] Employees who perceive bias react by downsizing contributions, disengagement, absenteeism, leaving, and withholding innovative ideas in greater percentages than those who do not perceive negative bias.[40] Gallup's State of the Workforce Report estimates active disengagement alone can cost around $500 billion per year.[41]

- Investing in retention is generally a better strategy than letting talent leave. The literature points to multiple practices to better manage daily work experiences.[42]
- Involve multiple leaders and employees in diversity management. It helps to hear concerns and possible solutions directly. Leadership engagement can supply needed visibility and even build relationships.
- Review how and to whom assignments that lead to visibility, networking, and promotions get assigned. Ultimately these experiences set up employees for their next opportunities.
- Ensure equal access to decision-makers. Personal biases of leaders are best influenced by proper engagement and access.[43] Establish specifically expressed corporate rules and consequences for unacceptable behaviors.
- Introduce senior-level advocates focused on careers of women and minorities, particularly if there are disparities. In a study by economists Sylvia Ann Hewlett, Melinda Marshall, and Laura Sherbin, this practice resulted in more inclusive leaders who created a "speak-up" culture, which decreased perceived bias up to 90 percent.[44] Addressing the underlying systems that keep inequality in place is important to changing the work environment.[45]
- Ensure inclusion-training incorporates easy-to-implement, skill-building tasks, and is ingrained with company goals and inclusive of targeted, positive approaches and messages that communicate acceptable behaviors.[46] As Kalev and Dobbin discuss in their essay in this volume, cultural inclusion training curricula can be particularly effective at teaching listening and

observational skills and increasing diversity and inclusion.[47] For example, affinity bias, the tendency to gravitate toward people similar to us, can be limited by requiring that hiring and promotion slates must include two or more qualified underrepresented candidates as well as two or more qualified women.[48]

Donofrio: This is an important topic. I have always understood that for business success, everyone is needed and welcomed and enabled to be their best. As a young engineer, I often wondered what we were missing based on who was not in the room where it happened. Clearly, there were very few women and even fewer people of color. How do we know that this all-white male group is going to give us the best answer? Are they really the only experts on this topic? While I struggled with these thoughts, I also better understood what needed to be done.

As a member of the National Action Council for Minorities in Engineering board, and later serving as its Chair, as a recipient of the Rodney D. Chipp Memorial Award from the Society of Women Engineers and eventually a member of their ranks, and as a recipient of the Renaissance Engineer Award from the Society of Hispanic Professional Engineers, as well as a frequent speaker at their national convention, I learned to use every opportunity to help turn bias into belief by simply enabling and amplifying the obvious. "She is right, just listen!" Who is not here, and where are they?

At IBM, we narrowed and focused our commitments in support of women and underrepresented minorities. We could not do everything and support everyone, but for those we could help, we would double down if not triple down our support and long-term commitment. This was not always the most popular corporate decision, given the natural bias to always keep your business commitment options open to change when and as required. But addressing long-term structural societal deficits requires much more than simply following wise and sage corporate advice. Once again, we had to make good business sense out of all of this, regardless of how emotionally connected our biases and beliefs were.

As I studied the processes around innovation between 2002 and 2008 for the U.S. Council on Competitiveness (see their National Innovation Initiative, which I helped lead and IBM strongly supported), all the pieces started to fall into place. Start with the problem and not the answer.[49] Enable an environment that supports open, collaborative, multidisciplined, global engagement. If you do not know who has the missing piece to the puzzle, why are you excluding people from engaging? Why let your bias determine the outcome? Innovation in business, government, education, for-profit, and nonprofit enterprises is the holy grail. Letting your bias control what you do is like starting with the answer. Every now and then, you may actually get it right. But more often than not, you will be wrong.

Perhaps it was easier for me to engage on these topics because I was often representing those who were not in the room to people who were in the room who looked

just like me. It is likely I had their biases if not beliefs working for me. Bias or belief. Right or wrong? Moral or immoral? If you are honest about what is really best for the enterprise now and in the future, the answer is obvious to and believable by all!

Sigur: In the 1980s, I took part in a "glass ceiling study" to assess whether there was indeed a barrier to promotion for women and minorities. The approach was generally straightforward, starting with an assessment of the numbers of employees of multiple demographics at each career stage. One of the products of the study was a series of pie charts showing percentages of diverse employees in the overall population, starting at entry levels and at successive levels of career advancement, later called the "Pac-Man series," as successive levels of career promotions showed smaller and smaller percentages of women and minorities who had made it through the ranks. The graphics, placed in the same location on successive pages, produced a flip-book animated illusion or movie when viewed in quick succession that was reminiscent of the chomping arcade character by Namco. The effect was dramatic and rallied support for addressing inequities. The white male population was "eating" women and minorities over time. Significant corrective actions included mentoring, leader training, implementation of representative promotion slates, and more consistent processes, with positive shifts in representation.

How do we establish clear practices to manage organizational demographics, such as hiring and retention? A diverse team starts with ensuring effective hiring.

- Establish objective criteria for each position.
- Ensure interviews are structured around skills-based questions to limit bias.
- Limit the use of referrals to avoid reaffirming social rather than objective-focused hiring.
- Supply explicit guidance on expectations for a "diverse candidate pool." Research shows that the odds of hiring or promoting a woman are seventy-nine times as great if at least two women are in the finalist pool, while the odds of hiring or promoting a nonwhite nominee are one hundred ninety-four times as great with at least two finalist minority applicants.[50]
- Avoid assessing candidates for "culture fit." Culture fit is frequently associated with shared interests, experiences, and backgrounds. When used as a selection criterion, culture fit often leads to homogeneity. Clear objective criteria and a common rubric to evaluate candidates help to avoid unintended impacts.[51]
- Place focus on recruiting diverse candidate employees. Although historically white alma maters of existing managers supply great references and those schools produce candidates of color, top Brown and Black talent should also be recruited from minority-serving institutions, including historically Black

colleges and universities (HBCUs), Hispanic-serving institutions (HSIs), Tribal colleges and universities (TCUs), and Asian American and Pacific Islander–serving institutions (AAPISIs).

Donofrio: Capitalizing on the belief that talent is equally distributed and the need for STEM talent is nearly endless, my colleagues and I focused on pathways for historically underrepresented communities in STEM. Included within the broad HBCUs are fifteen schools that meet the Certified Engineering requirements set by the Accreditation Board for Engineering and Technology. Our bias-turned-belief was that if we could ensure success for current students through industry internships and mentoring, and at the same time ensure success for faculty through industry and government funding, these colleges and universities with help and support could build out clear and strong pathways for local P–12 schools in their proximity. Focusing school by school, this initiative has started to take hold, built off this underlying belief that rich STEM talent is available locally and simply needs to be enabled. The added belief here is that constantly moving talent to opportunity may not always yield the fullest return on the investment. Moving opportunity to talent is a belief that counters the bias that talent must seek out opportunity and move toward it.[52]

Sigur: Leaders committed to developing diverse talent may take bold action. In response to learning that roughly three-quarters of the Black executive leadership in the corporation had undergraduate degrees from HBCUs, the CEO, Marilyn Hewson, allocated an annual investment for developing talent pathways produced by accredited minority-serving institutions (MSIs) with executive liaisons to develop programs to help both the students and the corporation. The multimillion-dollar investment resulted in multiple benefits: a dialogue on new and upcoming talent beyond the big name schools and frankly an effort to grow talent resident in not only HBCUs, but MSIs and community colleges serving other communities across the United States; partnerships with some of these schools and faculty on new contracts; and opportunities for the faculty of these schools to engage in big business engineering and technology—experiences most of them had not had.

While hiring talent is critical, retention and development must get equal attention. Data on executive leaders who are neither white nor male show that business practices for promotions and employee development should be reexamined. Generally, around 90 percent of the *Fortune* 500 CEOs are white men, usually justified by a lack of available and qualified candidates. In addition to pipeline challenges, research points to negative perceptions of companies with women and minority executive leadership by both outside stakeholders (such as equity investors and other CEOs) and internal stakeholders. Internal leaders experiencing reduced organizational identification following the appointment of a racial minority or female CEO (versus the appointment of a white male CEO) manifests in tendencies to supply less help in task guidance, less mentoring, and limited recommendations

for promoting minority colleagues.[53] Hopefully, these tendencies can be managed through the best practices captured by business scholars Michael L. McDonald, Gareth D. Keeves, and James D. Westphal.

- Ensure objective criteria are used in evaluations, with different metrics for skills, personality, and potential;
- Implement transparent promotion and talent-development review boards that understand company values and goals;
- Use structured and open mentoring; and
- Implement robust and transparent succession planning.[54]

Sigur: In my career, diversity initiatives have focused mostly on hiring, with some success in wider representation of both race and gender identities. But promotion candidates were mostly white and male. And their résumés and experiences supported them as the better candidates for advancement. Why were equally capable women and people of color becoming "less capable" over time?

An examination of contributing causes revealed that they were not given the "hard assignments": working challenges on the production floor, dealing with difficult customers, or meeting tight margin assignments. Anecdotally, some white women weren't being given "dirty work" in attempts to "protect" them; a Black woman wasn't being given an assignment because "she wouldn't like it"; and the Black men "might be too imposing." The result was that they missed assignments that would have developed needed skills because of stereotypes and biases. These issues were only revealed through one-on-one dialogues with supervisors, pointing out that these practices, while maybe well-intended, ensured that potential rising stars were being left behind. Without further prompting, the supervisors implemented immediate corrective action and our demographics improved.

Donofrio: Too often, we believe as leaders that offering ourselves as a mentor is going to consume us. How can I do any more than I am doing? How many protégés are too many? We approach this topic with a personal bias. Over the years, I have learned more from mentoring than I ever expected. The act of mentoring is simpler than I ever thought, and I learned to offer myself freely to anyone who would have me. People cringe when I say that. Yet how hard is it to listen and offer advice? How hard is it to connect one thing to another or one person to another? I tell my protégés to have many mentors. No mentor has all the answers to all your problems or questions. Keep the relationship real and lifelike. It is not about form; it is all about function. The temptation, and perhaps bias, is to document everything and report progress. To whom and for what reason? If you like what I have to say and it helps, you got what you wanted and needed and will come back as needed for more. Along the way, I will also learn through the very subtle process of reverse mentoring. I counsel, guide, and teach you while I also see and learn and better understand you. This is so critical as we embrace and work in a world that is increasingly more and more multigenerational. Baby

Boomers working side by side with Gen Zs, if not pre-Boomers working side by side with Gen Alphas. All those biases and beliefs within each generation need to be understood, heard, and hopefully reconciled. Generational biases run rampant. We all think and believe we know better about each other until we sit down and talk together to determine what the real issues and questions are so we can each contribute to the solution.

Sigur: Succession planning is recognized as a best practice to ensure the longevity and health of an organization. When we implemented the "who's in charge if you win the lottery and leave?" dialogue, it seemed like a good process; however, the result was an overwhelming number of nonminority male, albeit qualified, leaders ascending through the organizational ranks. Somehow we had institutionalized a glass ceiling! The leadership courage of the executive vice president of Lockheed Martin Space, Joanne Maguire, saw this and turned it around. Upon realizing the limitations of our leadership pipeline, she instituted mandatory succession planning meetings that started with a synopsis of each department's current and future plans and included a "wall-walk" of succession plans for key positions. Mandatory diverse slates were developed, to include both "ready now" candidates and candidates that could be ready in as many as fifteen years, including skill gaps and paths to promotion. Assignments were negotiated for every candidate to push them to their next level. The result was the most transparent talent-development effort I've seen and a massive shift in the promotion of qualified women, men, and underrepresented minorities to significant positions within the company.

Organizations should measure the impact of their actions against desired goals by keeping internal metrics. Just like other metrics of accomplishment, such as returns on investment or capital, the returns on training and other actions should be measured and expanded to include not only whether the actions took place and under what circumstances, but data on the effects for those the actions are intended to benefit. Data should be shared, as appropriate. Visibility supports accountability and the data inform decision-making. Increasing the numbers of minority groups represented in the workforce does not mean the company has embraced inclusivity or created a culture of belonging. Established objective criteria, scoring rubrics, and consistent practices provide information to assess effective performance. Regular climate surveys can measure progress and identify areas of concern.[55] In addition, as companies use analytical tools to support workforce decisions – from hiring and promotion to productivity and compensation – the risks of bias being introduced into these key processes need to be monitored and managed, as potentially toxic training datasets or nonrepresentative information may detrimentally influence and worsen existing conditions.[56]

Multiple tools exist to aid in the development of fair and equitable processes and in assessing progress. Examples include the tools developed by MITRE, available through the MITRE Social Justice Platform and the Aspen Institute, with mul-

tiple tools that cover a range of job quality attributes: wages, benefits, scheduling, legal rights, equity and inclusion, opportunity to build skills and advance, supportive work environment, and worker voice.[57]

When results can be driven by policy, leadership direction should be used to reduce bias, whether explicit or implicit. The character or culture of an organization reflects its leaders and their courage. It may be difficult to shift individual bias, but it is definitely possible to interrupt bad behaviors, at least in the near term.

Our collective stories reflect the positive actions of strong leaders who had influence. Because it was those strong leaders who, regardless of their personal biases, made the hard decisions to break barriers enabling "firsts," but just as important, introduced changes that inspired emerging leaders to engage the organization's stakeholders in enabling opportunities for sustainable change.

So, is it implicit bias or intentional belief? And if leadership has the courage of its convictions to transform the community it can impact, to shift away from systemic bias toward supporting equality, opportunity, and high values, does the answer matter?

ABOUT THE AUTHORS

Wanda A. Sigur is former Vice President of Lockheed Martin Civil Space and President and Founder of Lambent Engineering. She is an aerospace strategy and program management consultant for both emerging space exploration companies and traditional aerospace industry companies.

Nicholas M. Donofrio, a Fellow of the American Academy since 2005, is former Executive Vice President of Innovation and Technology at the IBM Corporation. He is the author of *If Nothing Changes, Nothing Changes* (with Michael DeMarco, 2022).

ENDNOTES

[1] Gallup, "Race Relations," https://news.gallup.com/poll/1687/race-relations.aspx (accessed May 2023).

[2] Gallup, "LGBTQ+ Rights," https://news.gallup.com/poll/1651/Gay-Lesbian-Rights.aspx (accessed July 30, 2023).

[3] Tessa E. S. Charlesworth and Mahzarin R. Banaji, "Patterns of Implicit and Explicit Attitudes: IV. Change and Stability from 2007 to 2020," *Psychological Science* 33 (9) (2022): 1347–1371.

[4] George B. Cunningham and Harper R. Cunningham, "Bias among Managers: Its Prevalence across a Decade and Comparison across Occupations," *Frontiers in Psychology* 13 (2022).

[5] "Businesses started caring a lot more about diversity after a series of high-profile lawsuits rocked the financial industry. . . . They have also expanded training and other diversity programs. But on balance, equality isn't improving in financial services or elsewhere. Although the proportion of managers at U.S. commercial banks who were Hispanic rose from 4.7 percent in 2003 to 5.7 percent in 2014, white women's representation dropped from 39 percent to 35 percent, and black men's from 2.5 percent to 2.3 percent. . . . Among all U.S. companies with 100 or more employees, the proportion of black men in management increased just slightly–from 3 percent to 3.3 percent–from 1985 to 2014. White women saw bigger gains from 1985 to 2000–rising from 22 percent to 29 percent of managers–but their numbers haven't budged since then." Frank Dobbin and Alexandra Kalev, "Why Diversity Programs Fail," *Harvard Business Review*, July–August 2016.

[6] Mona Mourshed, Ali Jaffer, Helen Cashman, et al., "Meeting the World's Midcareer Moment," Generation, July 2021, https://www.generation.org/wp-content/uploads/2021/07/Meeting-the-Worlds-Midcareer-Moment-July-2021.pdf.

[7] Knowledge at Wharton Staff, "Analyzing Effective Leaders: Why Extraverts Are Not Always the Most Successful Bosses," *Knowledge at Wharton*, November 23, 2010, https://knowledge.wharton.upenn.edu/article/analyzing-effective-leaders-why-extraverts-are-not-always-the-most-successful-bosses.

[8] Stuart W. Flint, Martin Čadek, Sonia C. Codreanu, et al., "Obesity Discrimination in the Recruitment Process: 'You're Not Hired!'" *Frontiers in Psychology* 7 (2016): 647.

[9] Inna Reddy Edara, "Anti-Asian Racism in the Shadow of COVID-19 in the U.S.A.: Reported Incidents, Psychological Implications, and Coping Resources," *Journal of Psychological Research* 2 (3) (2020): 13–22.

[10] Introduced in 1998, the Implicit Association Test is a computer-based test that "measures implicit attitudes by measuring their underlying automatic evaluation." Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz, "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test," *Journal of Personality and Social Psychology* 74 (6) (1998): 1464–1480. See also Charlesworth and Banaji, "Patterns of Implicit and Explicit Attitudes."

[11] Kirsten N. Morehouse and Mahzarin R. Banaji, "The Science of Implicit Race Bias: Evidence from the Implicit Association Test," *Dædalus* 153 (1) (Winter 2024): 21–50, https://www.amacad.org/publication/science-implicit-race-bias-evidence-implicit-association-test.

[12] American Psychological Association, "Implicit Bias," *APA Dictionary of Psychology*, https://www.apa.org/topics/implicit-bias (accessed May 1, 2023).

[13] Morehouse and Banaji, "The Science of Implicit Race Bias"; Kate A. Ratliff and Colin Tucker Smith, "The Implicit Association Test," *Dædalus* 153 (1) (Winter 2024): 51–64, https://www.amacad.org/publication/implicit-association-test; Rebecca C. Hetey, MarYam G. Hamedani, Hazel Rose Markus, and Jennifer L. Eberhardt, "'When the Cruiser Lights Come On": Using the Science of Bias & Culture to Combat Racial Disparities in Policing," *Dædalus* 153 (1) (Winter 2024): 123–150, https://www.amacad.org/publication/when-cruiser-lights-come-using-science-bias-culture-combat-racial-disparities-policing; and Alexandra Kalev and Frank Dobbin, "Retooling Career Systems to Fight Workplace Bias: Evidence from U.S. Corporations," *Dædalus* 153 (1) (Winter 2024): 213–230,

https://www.amacad.org/publication/retooling-career-systems-fight-workplace-bias-evidence-us-corporations.

[14] Calvin K. Lai and Mahzarin R. Banaji, "The Psychology of Implicit Intergroup Bias and the Prospect of Change," in *Difference without Domination: Pursuing Justice in Diverse Democracies*, ed. Danielle Allen and Rohini Somanathan (Chicago: University of Chicago Press, 2020); and Jack Glaser, "Disrupting the Effects of Implicit Bias: The Case of Discretion & Policing," *Dædalus* 153 (1) (Winter 2024): 151–173, https://www.amacad.org/publication/disrupting-effects-implicit-bias-case-discretion-policing.

[15] Jonathan Haidt and Selin Kesebir, "Morality," in *Handbook of Social Psychology*, 5th edition, ed. Susan T. Fiske, Daniel T. Gilbert, and Gardner Lindzey (Hoboken, N.J.: Wiley, 2010), 797–832.

[16] Manuel J. Galvan and B. Keith Payne, "Implicit Bias as a Cognitive Manifestation of Systemic Racism," *Dædalus* 153 (1) (Winter 2024): 106–122, https://www.amacad.org/publication/implicit-bias-cognitive-manifestation-systemic-racism.

[17] Hetey, Hamedani, Markus, and Eberhardt, "'When the Cruiser Lights Come On.'"

[18] Patrick S. Forscher, Calvin K. Lai, Jordan R. Axt, et al., "A Meta-Analysis of Procedures to Change Implicit Measures," *Journal of Personal Social Psychology* 117 (3) (2019): 522–559; and Chloe FitzGerald, Angela Martin, Delphine Berner, and Samia Hurst, "Interventions Designed to Reduce Implicit Prejudices and Implicit Stereotypes in Real World Contexts: A Systematic Review," *BMC Psychology* 7 (1) (2019): 1–12.

[19] Kalev and Dobbin, "Retooling Career Systems to Fight Workplace Bias."

[20] Lai and Banaji, "The Psychology of Implicit Intergroup Bias and the Prospect of Change."

[21] Cass R. Sunstein and Richard H. Thaler, "Libertarian Paternalism Is Not an Oxymoron," *The University of Chicago Law Review* 70 (4) (2003): 1159–1202.

[22] James MacGregor Burns, *Leadership* (New York: Harper & Row, 1978), 4.

[23] Bernard M. Bass, *Leadership and Performance beyond Expectations* (New York: Free Press, 1985), 27.

[24] Richard H. Thaler and Cass R. Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness* (New York: Penguin Group, 2009).

[25] Bruce J. Avolio, David A. Waldman, and Francis J. Yammarino, "The Four *I*'s of Transformational Leadership," *Journal of European Industrial Training* 15 (4) (1991): 9–16.

[26] Syed Talib Hussain, Shen Lei, Tayyaba Akram, et al., "Kurt Lewin's Change Model: A Critical Review of the Role of Leadership and Employee Involvement in Organizational Change," *Journal of Innovation & Knowledge* 3 (3) (2018): 123–127.

[27] Kalev and Dobbin, "Retooling Career Systems to Fight Workplace Bias."

[28] Fredson Kotamena, Pierre Senjaya, and Augustian Budi Prasetya, "A Literature Review: Is Transformational Leadership Elitist and Antidemocratic?" *International Journal of Sociology, Policy and Law* 1 (1) (2020).

[29] Edgar H. Schein, *Organizational Culture and Leadership*, 3rd ed. (Hoboken, N.J.: Wiley, 2004), 7.

[30] Ibid.

[31] Ibid., 223.

[32] National Commission on Innovation and Competitiveness Frontiers, *Competing in the Next Economy, The New Age of Innovation* (Washington, D.C.: Council on Competitiveness, 2020).

[33] National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, *Rising Above the Gathering Storm, Revisited: Rapidly Approaching Category 5* (Washington, D.C.: The National Academies Press, 2010).

[34] Bas Hofstra, Vivek V. Kulkarni, Sebastian Munoz-Najar Galvez, et al., "The Diversity–Innovation Paradox in Science," *Proceedings of the National Academy of Sciences* 117 (17) (2020).

[35] Sylvia Ann Hewlett, Melinda Marshall, and Laura Sherbin, "How Diversity Can Drive Innovation," *Harvard Business Review* 91 (12) (2013).

[36] Luis Emilio Gomez and Patrick Bernet, "Diversity Improves Performance and Outcomes," *Journal of the National Medical Association* 111 (4) (2019): 383–392.

[37] Committee on Homeland Security and Governmental Affairs, U.S. Senate, *Diversity Management: Expert-Identified Leading Practices and Agency Examples* (Washington, D.C.: United States Government Accountability Office, 2005).

[38] Nancy J. Cooke and Margaret L. Hilton, eds., *Enhancing the Effectiveness of Team Science* (Washington, D.C.: The National Academies Press, 2015).

[39] The methodology is based on a survey administered to a sample of 1,313 American workers, including Black, Hispanic/Latino, and Asian workers. Survey questions addressed the feelings, sources, and resulting impacts (actions) of the employees based on their experiences associated with unfair treatment due to their race or ethnicity. See Society for Human Resource Management, *SHRM Study: Racial Bias at Work Costs U.S. Employers Billions* (Trenton: New Jersey Industry and Business Association, 2021).

[40] Research was conducted using a survey (3,570 respondents: 1,605 men, 1,965 women; 374 Black, 2,258 white, 393 Asian, 395 Hispanic; ages twenty-one to sixty-five) and focus groups (both in-person [56 people] and virtual [250 people]). Analyses were performed by CTI (Center for Talent Innovation, currently known as Coqual). Data on employee potential were assessed using a methodology derived by CTI using an instrument to measure six attributes: ability, ambition, commitment, connections, emotional intelligence, and executive presence. Survey questions included a self-assessment of potential, and an assessment of how employees perceived their supervisors would assess their potential across all six fronts. Those that identified their own potential as greater than their supervisor's perception of their potential in two or more areas were identified as perceiving a negative bias. The percentages of employees who perceived negative bias ranged from 7.7 percent to 14.5 percent by groupings of race, sexual orientation, gender, immigrant/citizenship status, and ability/disability. A consortium of eighty-six human resource officers was used in developing the instrument. The focus of their study was to capture how bias manifests in corporations, how it impacts companies, and what can be done to disrupt it. See Sylvia Ann Hewlett, Ripa Rashid, and Laura Sherbin, *Disrupt Bias, Drive Value: A New Path Toward Diverse, Engaged, and Fulfilled Talent* (New York: Coqual, 2017).

[41] Ryan Pendell, "The World's $7.8 Trillion Workplace Problem," Gallup, June 14, 2022, https://www.gallup.com/workplace/393497/world-trillion-workplace-problem.aspx. In September 2023, the article was updated to reflect a new sum: $8.8 trillion.

[42] Stephanie J. Creary, Nancy Rothbard, and Jared Scruggs, "Improving Workplace Culture through Evidence-Based Diversity, Equity, and Inclusion Practices," PsyArXiv, July 1, 2021, https://doi.org/10.31234/osf.io/8zgt9.

[43] Adam D. Galinsky and Gordon B. Moskowitz, "Perspective-Taking: Decreasing Stereotype Expression, Stereotype Accessibility, and In-Group Favoritism," *Journal of Personality and Social Psychology* 78 (4) (2000): 708–724.

[44] Hewlett, Marshall, and Sherbin, "How Diversity Can Drive Innovation."

[45] Hetey, Hamedani, Markus, and Eberhardt, "'When the Cruiser Lights Come On.'"

[46] Creary, Rothbard, and Scruggs, "Improving Workplace Culture through Evidence-Based Diversity, Equity and Inclusion Practices."

[47] Kalev and Dobbin, "Retooling Career Systems to Fight Workplace Bias."

[48] Stefanie K. Johnson, David R. Hekman, and Elsa T. Chan, "If There's Only One Woman in Your Candidate Pool, There's Statistically No Chance She'll Be Hired," *Harvard Business Review*, April 1, 2016.

[49] National Commission on Innovation and Competitiveness Frontiers, *Competing in the Next Economy, The New Age of Innovation*.

[50] Ibid.

[51] Joan Williams and Sky Mihaylo, "How the Best Bosses Interrupt Bias on Their Teams," *Harvard Business Review*, November–December 2019.

[52] For additional information on aligning talent and opportunity, see Nicholas M. Donofrio, "Strengthen Innovation and Inclusion by Bringing Opportunity to Talent," *The Bridge: Linking Engineering and Society* 50 (5) (2021).

[53] "Organizational identification" is "a form of group identification in which an employee defines himself or herself in terms of involvement or membership in a particular corporation or other type of workplace. A growing body of evidence indicates that organizational identification underpins a range of important outcomes in work settings." For more about organizational identification, a term used to capture alignment of an individual and the organization's values and vision, see American Psychological Association, "Organizational Identification," *APA Dictionary of Psychology*, https://dictionary.apa.org/organizational-identification (accessed July 2023).

[54] Michael L. McDonald, Gareth D. Keeves, and James D. Westphal, "One Step Forward, One Step Back: White Male Top Manager Organizational Identification and Helping Behavior toward Other Executives Following the Appointment of a Female or Racial Minority CEO," *Academy of Management Journal* 61 (2) (2018): 405–439; and Jen Choi, Micha'el Monique Davis, Lilianna Kay Deveneau, et al., *Designing for Equity Starter Guide* (McLean, Va.: MITRE Social Justice Platform, 2022), https://sjp.mitre.org/insights/designing-for-equity-starter-guide.

[55] Marie A. Bernard, Frank Dobbin, Markus Brauer, et al., *Scientific Workforce Diversity Seminar Series (SWDSS) Seminar Proceedings: Is Implicit Bias Training Effective?* (Washington, D.C.: National Institutes of Health, 2021), https://diversity.nih.gov/sites/default/files/media-files/documents/NIH_COSWD_SWDSS_Implicit_Bias_Proceedings_508.pdf.

[56] Mike Capps, Bob Darin, Nuala O'Connor, et al., *Algorithmic Bias Safeguards for Workforce* (New York: Data & Trust Alliance, 2023), https://dataandtrustalliance.org/Algorithmic_Bias_Safeguards_for_Workforce_Overview.pdf.

[57] "A Framework for Assessing Equity in Federal Programs and Policies," MITRE Social Justice Platform, https://sjp.mitre.org/insights/60f1e225b1d934001a56df4b (accessed December 5, 2023); "A Systems Analysis of the Black-White Racial Wealth Gap in the District of Columbia" MITRE Social Justice Platform, https://sjp.mitre.org/racial-wealth-gap (accessed December 5, 2023); Jonathan Rotner, *An Equity Guide for Techies* (McLean, Va.: MITRE Social Justice Platform, 2022), https://sjp.mitre.org/insights/Equity%20Guide%20for%20Techies; Choi, Davis, Deveneau, et al., *Designing for Equity Starter Guide*; and The Aspen Institute, "Tools: Equity and Inclusion," last modified September 21, 2023, https://www.aspeninstitute.org/longform/job-quality-tools-library/section-4-strengthening-practices-to-improve-job-quality/tools-equity-and-inclusion.

Mirror, Mirror, on the Wall, Who's the Fairest of Them All?

Alice Xiang

Debates in AI ethics often hinge on comparisons between AI and humans: which is more beneficial, which is more harmful, which is more biased, the human or the machine? These questions, however, are a red herring. They ignore what is most interesting and important about AI ethics: AI is a mirror. If a person standing in front of a mirror asked you, "Who is more beautiful, me or the person in the mirror?" the question would seem ridiculous. Sure, depending on the angle, lighting, and personal preferences of the beholder, the person or their reflection might appear more beautiful, but the question is moot. AI reflects patterns in our society, just and unjust, and the worldviews of its human creators, fair or biased. The question then is not which is fairer, the human or the machine, but what can we learn from this reflection of our society and how can we make AI fairer? This essay discusses the challenges to developing fairer AI, and how they stem from this reflective property.

How can we develop fairer artificial intelligence (AI) that does not reflect, entrench, and amplify societal biases? There are three major categories of interventions: data curation, algorithmic methods, and policies around appropriate use. The first is motivated by the fact that AI, like a mirror, tends to reflect the biased patterns present in its training data. If a voice recognition model is trained predominantly on audiobooks, it might learn how to accurately understand "standard" varieties of language but struggle to understand accents, dialects, or speech impediments.[1] In domains like computer vision and speech language technologies, diversity in the appearance and voices of the individuals represented in the training data is key to avoid the creation of biased AI models.[2] The second set of approaches is algorithmic interventions. AI is not a perfect mirror, so by imposing constraints or changing the objectives for the model's optimization, algorithmic fairness practitioners seek to warp the mirror, making the outputs more accurate or fairer. Much of the literature in this space focuses on defining specific fairness metrics and developing preprocessing, in-processing, or postprocessing methods to make the model's outputs perform better on the basis of those fairness metrics.[3] The third set of approaches focuses on defining when AI or humans should be used. For example, the moratoriums several U.S.

https://doi.org/10.1162/daed_a_02058

jurisdictions put in place around law enforcement's use of facial recognition and the European Union's AI Act, which prohibits certain high-risk categories of AI, fall under this category.[4] Combinations of these approaches are vital to addressing bias mitigation, but as this essay will discuss, there are many technical, legal, and operational challenges to creating fairer AI in practice.

Starting first with data curation, regarding AI as a mirror implies looking beyond the AI model itself to the societal context surrounding it, which it reflects in turn. Just as a parent can shape their child's worldview by controlling the information and experiences the child is exposed to, AI developers can similarly mold their AI through their data selection. Most image datasets are sourced exclusively from a few developed countries.[5] Biases in computer vision models have largely been attributed to a lack of sufficient representation of women and minorities in such datasets.[6] Like humans who find it easier to accurately distinguish people in the majority ethnic group they grew up in, human-centric computer vision models tend to more accurately recognize the types of people featured in their training data.[7] Moreover, lack of diversity in the background, objects, clothing, and other features can lead to additional biases. For example, what does "soap" look like? The answer can differ depending on which part of the world you are from. Researchers found that object detection algorithms trained predominantly on data from higher-income countries struggled to accurately recognize objects in lower-income countries.[8]

The digital divide can further exacerbate inequities in whose interests are reflected in datasets. For example, in 2012, Boston-based startup Connected Bits launched its StreetBump app, leveraging accelerometer and GPS data to automatically detect potholes and inform the city where to direct resources to fix them.[9] The data collected by the app, however, painted a distorted picture of the prevalence of potholes in the city.[10] People in lower-income neighborhoods were less likely to have smartphones to download the app, leading to systematic underrepresentation of the number of potholes in their neighborhoods needing repair.

There are thus strong normative arguments for collecting carefully curated, large, diverse, representative datasets to tackle algorithmic bias. While this statement has become a truism in the algorithmic fairness community, how to collect such datasets in practice is an unsolved problem.[11] Much of the progress in the past few decades of AI development has stemmed directly from questionable data-collection practices. In the early days of computer vision, images were sourced in highly controlled bespoke settings.[12] Researchers would set up photography studios to take pictures of subjects. These image datasets were consequently very small and highly constrained. The poses, backgrounds, and demographics of the people represented in these datasets were greatly limited. Computer vision and AI more generally were revolutionized by the development of large, publicly available web-scraped datasets. ImageNet, consisting of fourteen million images

scraped from the internet with annotations for the objects in the images, was revolutionary in enabling computer vision scientists to train their models on much larger-scale data than was previously possible.[13] In the ImageNet Large Scale Visual Recognition Challenge, AlexNet (a convolutional neural network) won, substantially beating the runner-up. A key feature of AlexNet was the depth of its network, which relied on a large training dataset. The success of AlexNet, one of the most influential developments in computer vision, contributed to the explosion in deep learning.[14]

While ImageNet and AlexNet were tremendously beneficial for the acceleration of AI development, they also set AI developers along a path that depended on vast amounts of data and computation. Achieving state-of-the-art models required large corpora of data that could not easily be obtained through curated, bespoke methods. Web-scraping, the method by which ImageNet was created, became the norm, with most large datasets since ImageNet relying on that method.[15] While web-scraping large amounts of online data leveled the playing field to some extent, it also carried with it significant ethical challenges. In recent years, ImageNet has faced criticism for its lack of informed consent, offensive labels, and problematic images, all of which are artifacts of its collection methodology.[16]

This dependency on web-scraped images has carried over to algorithmic fairness efforts. My recent work has discussed this issue in depth, exploring the tensions that emerge between fairness and privacy in operationalizing data-collection efforts for human-centric computer vision.[17] For example, in 2018, IBM released the Diversity in Faces (DiF) dataset.[18] Like most large-scale computer vision datasets at the time, this dataset was based on images scraped from Flickr with permissive licenses. IBM's contribution was to find a diverse subset of face images and provide labels of relevant features, enabling the dataset to be used by fairness researchers checking for biases in their models. Even though ImageNet, COCO (Common Objects in Context, another large web-scraped image dataset), and other major datasets similarly featuring Flickr images with humans had been available for years without any lawsuits, the launch of DiF was immediately fraught. Not only was IBM sued under the Illinois Biometric Information Privacy Act (BIPA) for processing individuals' biometric information without appropriate informed consent, but Microsoft and Google were also sued as downstream users of the dataset.[19] DiF was immediately removed by IBM, and not long afterward, IBM announced that it would be pulling away from facial recognition technologies in general.[20] Notably, other Flickr-based computer vision datasets remain available and have not faced any lawsuits. In 2021, ImageNet creators voluntarily decided to obscure the faces of image subjects (note that their bodies are not otherwise obscured), but COCO remains available without any obfuscation of faces or bodies.[21] Taking the DiF dataset as a starting point, let us consider the minefield of constructing a "fair" human image dataset.

For simplicity, I will consider "fair" to simply mean a dataset that is legally compliant, as globally diverse and free of biases as possible, and large and realistic enough to develop a state-of-the-art model. The benefits of web-scraped datasets are that they are large and realistic. That is not to say that they are free from biases. In fact, they tend to exhibit biases reflective of the platform aggregating the data and of society as a whole. For example, studies have shown that images on Flickr tend to be biased toward Western developed countries, where most of Flickr's users are located.[22] In addition, AI trained on such datasets tend to learn stereotypical patterns, such as associating women with domestic spheres and men with public spheres.[23] For instance, commonly used visual recognition datasets feature women cooking far more often than men cooking, teaching AI models to associate women with the activity of cooking.[24] Similar stereotypical trends have been found in word embeddings from large text corpora that can be used to train language models.[25]

Moreover, as the DiF authors discovered, web-scraping presents many legal issues. In the past few years, lawsuits stemming from U.S. state biometric information privacy laws and the European Union's General Data Protection Regulation have raised awareness that using face images for AI without informed consent is inappropriate from a legal compliance perspective. Facebook reached a landmark settlement in 2021 for $650 million in a BIPA lawsuit contesting their processing of users' biometric data through facial recognition technology to support their automated tag-suggestion feature.[26] Nonetheless, researchers in computer vision still rely heavily on such datasets. Many do not see any alternative for how they could conduct research in this field otherwise.[27] Indeed, the recent explosion in generative AI technologies has only further exacerbated this issue, requiring even larger amounts of data to train, and normalizing the idea that participation at the frontier of AI necessitates training such models indiscriminately on content from across the internet.

Copyright has also increasingly become a concern in data curation. For models trained on large web-scraped corpora of text, images, and video, rarely is the permission of content creators sought prior to using their content for AI development. This has especially presented problems in the generative AI context, where AI models not only benefit from the use of copyrighted content but often generate new content inspired by such inputs, but without appropriate attribution. Creators of web-scraped computer vision datasets have historically emphasized their reliance on data with permissive licensing as an argument for why they are not infringing on IP rights.[28] While such arguments might have sufficed when copyrighted materials were used to train AI models for tasks unrelated to creative pursuits – such as transcribing text or drawing bounding boxes, key points, or segmentation masks on humans and objects – generative AI presents new concerns. If an artist uploads their work to a public platform with a permissive license for the content to be redistributed, have they

also agreed to allow the generation of derivative works that might imitate their style or content, possibly to the extent of cannibalizing their business? Arguably, until recently, few artists could have foreseen such consequences.

In addition, collecting globally representative data presents many practical complexities due to real-world geopolitical divisions. Contracting with and obtaining informed consent from people around the world is challenging given differences in local laws, cultures, and languages. Privacy and intellectual property rights vary substantially across jurisdictions. Many countries also have data localization laws that erect barriers to the transfer of their residents' data outside of their country.[29] Economic sanctions can further affect the extent to which some countries can be represented in AI datasets. These constraints add a geopolitical dynamic to what AI models learn about the world. Similar to how China's "Great Firewall" has led to a distinct internet experience for Chinese netizens, legal and political barriers around data collection can lead to more fragmented AI development, with parochial AI models that primarily only understand the people and patterns in their own geographies.[30]

Beyond web-scraping, another approach to assembling large, diverse datasets is to use existing repositories of stock images taken by professional photographers. While this is less problematic than the web-scraping approach given that photographers and the image subjects were possibly compensated for their work, it is difficult to say whether these individuals could have anticipated that their works would be used to develop AI. Especially in an era of generative AI, allowing your photos to be used to train AI could have downstream implications, such as content featuring your likeness (if you are the image subject) or artistic style (if you are the photographer) being generated by the AI in response to prompts you have no control over. This is especially problematic if the generated content is misleading or offensive. Moreover, stock photos taken by professional photographers look very different from the more naturalistic images that AI is likely to encounter in deployment. The lighting is often perfected, the setting and poses staged, and the image subjects more conventionally attractive. AI developed using such images can have a harder time recognizing people or objects in the real world due to this domain shift, or difference in the distributions of the training data and deployment context data.[31]

Bespoke data collection that reflects the deployment context is thus generally the best approach in that it enables more control over the data-collection process and assurance that both artists and subjects are fully informed about how their data are likely to be used. Operationalizing bespoke data collection, however, is very difficult. It requires developing business relationships with people around the world who can contribute to data-collection efforts. As a result, many companies specialize specifically in data collection. They recruit large numbers of crowd-workers from around the world to perform various data collection and an-

notation tasks for client companies. These companies have faced significant scrutiny, however, as some have deployed problematic recruitment practices to source diverse crowds and others have failed to provide appropriate employment conditions for data annotators.[32] Collecting billions of images through such methods can also be cost-prohibitive for smaller companies and researchers.[33]

Bespoke data collection further presents the challenge of requiring diversity specifications. In an underspecified dataset for which demographic balance is required of only a few attributes, like gender, age, and ethnicity, the images tend to look highly staged and homogenous.[34] It is easiest for people to take pictures of themselves standing or sitting, facing the camera, inside or right outside of their home. Additional requirements, such as a variety of poses, backgrounds, lighting conditions, number of people/objects, and interactions between them, can exponentially increase the complexities of data collection. This is especially the case given that it is not enough to provide a generic specification that diversity along these dimensions should be maximized. Checking for and ensuring diversity requires annotations specifying what pose, background, and lighting conditions are featured in the image. This requires a taxonomy for such attributes and extensive time and resourcing for image subjects or annotators to label the images. How do you adequately define the parameters for how the "real world" looks?

Moreover, the annotations related to the diversity of image subjects themselves can be highly contentious. For example, there are often concerns around bias associated with race, ancestry, or ethnicity, but collecting data on these attributes to check for bias can be complex given the social construction of such attributes. Different countries vary widely in how their census surveys characterize relevant ethnic groups, with some even refusing to collect race data, so there is no singular taxonomy that is consistent across the world.[35] Even the act of asking someone for these attributes can raise privacy concerns given the sensitive nature of such data, along with worries that the data will be used to discriminate against them (rather than prevent discrimination).[36]

Given these challenges with real data, there has been growing interest in the potential for synthetic data, leveraging recent advances in generative AI. Bypassing the need for real people, synthetic data can reduce many of the legal challenges with using real data, but issues of fairness persist. Creating synthetic data is like creating a microcosm of the world: while developers might be freed from some of the constraints of reality, that freedom also creates more room for subjectivity. For example, every conceivable skin tone, nose/face/eye shape, hairstyle, or body type is theoretically possible to generate synthetically, but with this flexibility comes more need for developers to specify the parameters of interest. It is like asking an artist to draw a fully inclusive representation of humankind. Biases and limitations of the artist's imagination can translate into a narrower worldview compared with large amounts of real-world data. For example, an artist might draw figures of vary-

ing skin tones and facial features, but all with similar body types and clothing styles, with backgrounds and objects that reflect a middle-class American living standard. Like a parent raising their child in a virtual simulation, AI developers who rely on synthetic data theoretically have more control over what their "child" is exposed to, but it can be difficult to create a synthetic environment as rich as reality but lacking the biases of the real world. For AI that operates exclusively in a synthetic environment, like AI avatars in video games, such a domain shift is not necessarily a problem. In most cases where the AI interacts with the real world, however, algorithmic bias is relevant, and this difference between the "world" where the AI is developed versus deployed can exacerbate potential biases.

Addressing data diversity and sourcing, however, is only the first part of the problem. Having a globally representative dataset simply ensures that the mirror is not warped, and your model reflects a more accurate representation of the world. The reflection we see in a perfect mirror is nonetheless often not flattering. Societal inequities and injustices that are present in the real world will naturally be reflected in such data. This presents one of the major challenges of algorithmic fairness: how to conceptualize a fair society and enable our AI models to promote rather than work against such a conception.

Early work highlighted the challenge of optimizing for multiple fairness definitions simultaneously. Researchers quickly proved impossibility theorems showing that some of the common fairness metrics conflicted with each other. Specifically, a model could not simultaneously be well-calibrated and have equalized odds across demographic groups if the demographic groups had different baselines.[37] The impossibility theorem inspired greater technical interest in the problem of algorithmic fairness.[38]

While the idea that data might reflect problematic patterns is increasingly accepted, the question of how to address these patterns is much less clear. While the algorithmic fairness literature features many solutions that imply differing thresholds or quotas for various sensitive attribute groups (that is, attributes receiving special legal protections like race, gender, or age), such solutions could be highly suspect viewed through a legal lens. As scholars have recently highlighted, it might not seem immediately evident that Supreme Court deliberations on affirmative action in higher education might have any bearing on algorithmic fairness.[39] But there are strong parallels that imply that if there were federal antidiscrimination litigation around algorithmic bias mitigation, many of the proposed methods could be deemed illegal.

In recent decades, the Supreme Court has increasingly turned toward anticlassification doctrine in its rulings.[40] Anticlassification is akin to colorblindness and implies that the fundamental goal of antidiscrimination law should be to prevent differential classification or treatment of individuals based on their protected

attributes. This contrasts with antisubordination, the doctrine that holds that antidiscrimination law should seek to actively dismantle historical discriminatory structures. Lyndon Johnson famously articulated the antisubordination underpinnings of affirmative action during his 1965 commencement address at Howard University: "You do not take a man who for years has been hobbled by chains, liberate him, bring him to the starting line of a race, saying, 'You are free to compete with all the others,' and still justly believe you have been completely fair."[41]

While debates about affirmative action have been active and controversial for many decades, the algorithmic fairness context highlights unique dimensions.[42] Cases like *Bakke*, *Grutter*, and *Gratz* conveyed the message to schools that affirmative action is only permissible if it cannot be easily quantified.[43] Quotas and point systems were patently unconstitutional, whereas holistic systems that used race as one of many factors were permissible. These types of decisions provided actionable guidance for human admissions officers who could keep an eye on racial composition of the class when making decisions, without ever formally quantifying any affirmative action boost. Such obfuscation is much more difficult for an algorithm.[44]

But the recent *Students for Fair Admissions* joint decision closed off even these approaches, solidifying the court's adoption of the anticlassification stance, as it struck down the affirmative action programs at Harvard and the University of North Carolina.[45] On the one hand, the court faulted these universities for their failure to provide quantifiable metrics for success (such as how much diversity is sufficient to obtain their educational objectives). But on the other hand, the court found their programs to be unconstitutional for the implicit quotas they adopted: for Harvard, "how the breakdown of the class compares to the prior year in terms of racial identities," and for the University of North Carolina, whether the "percentage enrollment within the undergraduate student body is lower than their percentage within the general population in North Carolina."[46] The court also declared that race could never be used as a negative factor, which in the zero-sum game of college admissions, implied that race could not be considered directly as a factor.[47] The only allowance the court gave to schools was that they could consider based on applicants' essays the possible impact of race on their experiences, provided that such essays highlighted the applicants' courage, determination, or other positive attributes.[48] The court has thus left very little room for explicit race-conscious antidiscrimination interventions, potentially posing challenges for the algorithmic fairness community, whose work typically involves formalizing a fairness metric, constraint, or objective that is conscious of the protected attribute, with the goal of affirmatively changing the model to be "fairer."[49]

The technical formalism of AI ethics, however, can also be used to reframe these contentious societal debates with greater clarity. At a time when it is common to bemoan the bias, opacity, and lack of accountability of AI, which is increasingly used throughout our society, does it make sense to incentivize either ignorance or

obfuscation of biases in such technology? AI developers seeking technologies that do not perpetuate societal biases already encounter many challenges to even testing for bias. As discussed above, privacy laws strongly disincentivize the collection of sensitive attribute data that is necessary for conducting bias audits. Should legal doctrine in antidiscrimination law further disincentivize developers from taking any action once bias has been discovered, out of fear of being successfully sued for (reverse) discrimination?

Understanding the connection between algorithmic fairness and broader societal debates about equity thus raises the stakes of these debates. Not only are courts debating the admissions criteria for elite schools, but such legal decisions codify normative principles that can influence the extent to which developers are legally allowed to modify increasingly ubiquitous algorithms to avoid amplifying bias against people from marginalized communities, despite their sway over decisions around recidivism, employment, credit, or other high-stakes domains. In other words, there may be a limit to how much developers can do to reduce the harm done by their own work.

At the heart of such societal debates is the tension between erasing versus mitigating the effects of systemic discrimination. Outside of the algorithmic context, proponents of anticlassification would argue that the goal should be to desensitize people to sensitive attributes like race and pursue a colorblind society.[50] Algorithmic fairness questions this notion given that AI trained without features like race are by default colorblind yet can still be racist. The richness of big data implies the presence of proxy variables and patterns correlated with race and other sensitive attributes that can be learned by a model that is not explicitly given sensitive attribute data.[51] For example, in the famous COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) case, the algorithm did not have any direct information about the race of the defendants. Nevertheless, because the training data reflected broader national trends whereby Black defendants had higher rates of re-arrest, likely due to disproportionate policing practices, the COMPAS algorithm leveraged features correlated to race, such that Black individuals were more likely to be incorrectly labeled as having high-recidivism risk.[52] Unlearning or avoiding biases on an algorithmic level typically requires knowledge of sensitive attribute information.[53] Algorithmic fairness thus errs on the side of antisubordination. While humans might be able to argue that ignoring race is an effective way to address racism, this is much more difficult for AI. Teaching AI the notion that racism is bad and should be avoided requires providing models some data about race.

Another way to view this debate through the lens of algorithmic fairness is to consider statistician George Box's famous quote, "All models are wrong, but some are useful."[54] Although he was addressing statistical models more generally rather than AI models, let alone bias in AI models, his insights still ap-

ply. Opponents of algorithmic bias mitigation efforts often resist interventions that are motivated by social justice inclinations out of concern that they are tampering with what is correct, true, or accurate. Indeed, the fact that fairness and accuracy in AI are often framed in terms of a trade-off is reflective of this idea.[55] The reality, however, is that all AI models are approximations of reality as conveyed to them via the data they are trained on. They are approximations built upon approximations of reality, and thus riddled with inaccuracies. For example, researchers found bias in health care algorithms that used cost of care as a proxy for health care need.[56] The training data reflected the pattern that Black patients of similar sickness levels to white patients receive less health care, so the model learned to downgrade the risk level of Black patients. Acknowledging these imperfections, the question then is how should we correct them? Bias mitigation efforts, instead of being framed as introducing additional inaccuracies, should be viewed as correcting existing inaccuracies in a direction that is more favorable from an equity perspective.[57]

This distinction can also be framed as a difference between prediction versus decision-making. Is the goal to have a mirror that as accurately as possible reflects reality in order to make accurate predictions? Or is the goal to improve upon the world and make it fairer? If AI is used purely for predictive tasks, like predicting whether someone will be re-arrested, then bias mitigation is less relevant given that reflecting societal biases accurately is helpful for making accurate predictions. But if AI is used for decision-making, there is a normative element that implies a need for bias mitigation. For example, deciding who should be denied bail is different from predicting who might be re-arrested. Many harms related to AI ethics stem from the conflation of a normative task with a descriptive one. If the goal is to decide who should be detained because they are more likely to commit a crime, then it is important to separate the bias of over-policing from the ground truth of crimes committed. This separation process is precisely what bias mitigation should aim to do.

The rise of generative AI technologies might further bring these debates into the content generation sphere. What content should be considered biased or discriminatory? These questions have long challenged content platforms, which typically rely on a combination of community guidelines, automated flagging of objectionable content, user reports, and human content moderators. Such efforts have thrown content platforms into contentious societal debates around whether their efforts are reasonable corrections to avoid disinformation versus problematic distortions of free speech.[58] With generated content from AI, however, the debate shifts: the question is no longer what human content is permissible to be shared on a platform, but rather what AI content should be generated. If an image generator consistently generates images of men whenever prompted with terms like "CEO," "intellectual," or "director," the AI might entrench existing societal

stereotypes.[59] To the extent such AI-generated images are then used to make or inspire art, movies, or media, they will amplify these biases.

On the one hand, from an antisubordination standpoint, this should create a responsibility on the part of the AI developer to take active measures to ensure the content generated reflects a less-biased view of the world. On the other hand, given current controversies around content moderation policies, it is likely that such affirmative efforts to create more balanced representations will be politically fraught.

In light of all of these challenges to implementing bias mitigation in practice, it is worth addressing the skepticism as to whether such efforts should even be pursued. Any fairness efforts predicated on having access to diverse data or sensitive attribute data necessitates the collection of yet more data, often about people from vulnerable, underrepresented populations, creating potential tradeoffs between fairness and other values like privacy.[60] Any attempts to rebalance the benefits or harms of algorithmic systems across demographic groups might cause significant political controversy, as the previous section discussed. It is tempting for such debates to go to extremes – for example, concluding that privacy must be protected at all costs – so fairness efforts requiring the collection of sensitive information at scale should be immediately halted.

For instance, some scholars have highlighted the concept of "horizontal relationality" in privacy, whereby the disclosure of private information by one individual could affect another individual's privacy, particularly in the context of machine learning and AI.[61] An example they use is if someone shares an image of their tattoo for a tattoo recognition model, the inclusion of their tattoo image in the training data for the tattoo recognition model could affect the model's ability to recognize similar tattoos on other individuals. If the tattoo recognition model is used by law enforcement to identify potential suspects, this could impact those individuals' privacy.

While horizontal relationality has been primarily characterized in a negative light – how one person's sacrifice of their privacy can force others to sacrifice their privacy – what's lost in this discussion is that there are benefits to horizontal relationality as well. In particular, such analyses assume an antagonistic relationship with technology, in which the goal is to ensure that AI does not work well for you. Such an attitude is typically motivated by concerns around AI surveillance: that AI primarily is being deployed by governments, employers, and others with power to surveil and deprive lower-powered individuals of autonomy and self-determination.[62] Many AI applications, however, lack this antagonistic relationship. Individuals buying a camera with AI autofocus for use in taking personal photos generally want the camera to be able to focus on their faces, their family's faces, and their friends' faces as accurately as possible. There is not necessarily a surveillance risk if the individual is taking photos and storing them on their drive

for personal consumption. While theoretically the person sharing images publicly on social media might create some surveillance risk for the individuals in the photos, that is unrelated to the functionality of the AI autofocus. If the autofocus worked poorly, the individual would likely just spend more time trying to get a good shot rather than give up entirely on sharing their lives on social media.

Even in high-stakes scenarios like law enforcement use of facial recognition to find suspects, it is unclear that reducing the performance of the AI model provides any benefits from the perspective of reducing surveillance. Much of the outcry against such high-risk use cases stem precisely from the negative impacts of poor performance of such models. In particular, there have been several notable cases in the United States of Black men being wrongfully arrested due to faulty facial recognition matches.[63] The question of whether law enforcement use of facial recognition is acceptable (a topic beyond the scope of this essay) is distinct from the question of whether better or worse accuracy of technologies is preferable.

If we assume such technologies will continue to be in use, then better accuracy benefits everyone other than those trying to evade law enforcement. Misrecognition for individuals with less societal privilege is especially pernicious since these individuals are less likely to have access to recourse to prove the mistake. This could include access to effective legal counsel, knowledge of relevant legal protections, and funds needed for bail. But weakening such surveillance technologies or making them less accurate won't benefit those people most harmed now; it will simply make such wrongful arrests *more* likely. So, if such surveillance technologies are in widespread use, there is a strong argument for maximizing the accuracy of such technologies for all groups, with the greatest benefits to those who are most victimized by the errors. More generally, regardless of whether a technology is low or high risk, trying to combat biases in the technology is a worthy endeavor. Critiques about algorithmic fairness efforts would more accurately be framed as critiques of specific AI use cases, especially ones that are surveillance oriented.

Separating these two considerations – whether a technology should be banned versus whether it should be improved – is of critical importance when attempting to operationalize algorithmic fairness. Why should humanity not have its cake and eat it too? While there might be some technologies so dangerous that outright bans are the only morally permissible response, the majority of AI technologies fall into a gray area in which their use should be conditional on appropriate safeguards. Navigating such gray areas requires taking action to address issues of bias, even when doing so requires carefully balancing other ethical desiderata.

Compared to other forms of technology, a distinguishing feature of AI is its capacity to learn from the data presented to it. This learning process transforms AI from a purely objective, rational machine to a mirror reflecting a version of our world. What makes AI ethics a fascinating discipline is that the

problems in this subfield are a microcosm for broader societal problems. The key difference, however, is that AI is our own creation, which sets a stronger moral requirement for us to address these problems and avoid employing AI that perpetuates and entrenches existing societal problems. Moreover, in certain ways, we have more control over AI models than we do over broader society. For example, although collecting a globally diverse training dataset to train a facial recognition model is extremely difficult, it is still easier than counteracting the biases of billions of peoples' human facial recognition. Thus, developing fairer AI is a difficult task, not simply because AI is often a black box, but also because AI reflects society and all its complexities. AI developers are often faced with difficult unsolved ethical questions that cut to the core of contemporary debates: What should you do to rectify historical injustices? How can you achieve fairness or diversity? To address these questions, we must think of AI not as a separate entity, a jumble of numbers and code, but rather as a mirror reflecting our society.

ABOUT THE AUTHOR

Alice Xiang is Global Head of AI Ethics at Sony Group Corporation and Lead Research Scientist at Sony AI. She has published in such journals as the *Harvard Journal of Law & Technology*, *Tennessee Law Review*, *University of Chicago Law Review*, and *Yale Law Journal* and in publications from machine learning conferences like *NeurIPS*, *ICML*, *ICLR*, *ICCV*, and *FAccT*.

ENDNOTES

[1] Orestis Papakyriakopoulos, Anna Seo Gyeong Choi, Jerone Andrews, et al., "Augmented Datasheets for Speech Datasets and Ethical Decision-Making," paper presented at the ACM Conference on Fairness, Accountability & Transparency, Chicago, Ill., June 12, 2023, https://arxiv.org/abs/2305.04672.

[2] Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research* 81 (2018): 1–15, http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

[3] B. d'Alessandro, Cathy O'Neil, and Tom LaGatta, "Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification," *Big Data* 5 (2) (2017): 120–134.

[4] Electronic Privacy Information Center, "State Facial Recognition Policy," https://epic.org/state-policy/facialrecognition (accessed December 12, 2023); and European Commission, *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts* (Luxembourg: Publications Office of the European Union, 2021), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

[5] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten, "Does Object Recognition Work for Everyone?" paper presented at the Institute of Electrical and Electronics Engineers/Computer Vision Foundation Conference on Computer Vision and Pattern Recognition, Workshop 52, Long Beach, Calif., June 16, 2019, https://openaccess.thecvf.com/content_CVPRW_2019/html/cv4gc/de_Vries_Does_Object_Recognition_Work_for_Everyone_CVPRW_2019_paper.html; and Keziah Naggita, Julienne LaChance, and Alice Xiang, "Flickr Africa: Examining Geo-Diversity in Large-Scale, Humancentric Visual Data," paper presented at the Association for the Advancement of Artificial Intelligence/Association for Computing Machinery Conference on AI, Ethics and Society, Montreal, August 8, 2023.

[6] Buolamwini and Gebru, "Gender Shades."

[7] P. Jonathon Phillips, Fang Jiang, Abhijit Narvekar, et al., "An Other-Race Effect for Face Recognition Algorithms," *ACM Transactions on Applied Perception* 8 (2) (2011): 1–11, https://doi.org/10.1145/1870076.1870082; and Elinor McKone, Lulu Wan, Madeleine Pidcock, et al., "A Critical Period for Faces: Other-Race Face Recognition Is Improved by Childhood But Not Adult Social Contact, Scientific Reports," *Nature*: *Scientific Reports* 9 (2019), https://www.nature.com/articles/s41598-019-49202-0.

[8] DeVries, Misra, Wang, et al., "Does Object Recognition Work for Everyone?"

[9] Connected Bits, "Street Bump," https://connectedbits.com/street_bump (accessed December 12, 2023).

[10] Kate Crawford, "The Hidden Biases in Big Data," *Harvard Business Review*, April 1, 2013, https://hbr.org/2013/04/the-hidden-biases-in-big-data.

[11] Weixin Liang, Girmaw Abebe Tadesse, Daniel Ho, et al., "Advances, Challenges, and Opportunities in Creating Data for Trustworthy AI," *Nature Machine Intelligence* 4 (2022): 669–677, https://doi.org/10.1038/s42256-022-00516-1; and Alice Xiang, "Being 'Seen' vs. 'Mis-Seen': Tensions between Privacy and Fairness in Computer Vision," *Harvard Journal of Law & Technology* 36 (1) (2022): 1–60, https://doi.org/10.2139/ssrn.4068921.

[12] Inioluwa Deborah Raji and Genevieve Fried, "About Face: A Survey of Facial Recognition Evaluation," paper presented at the Association for the Advancement of Artificial Intelligence 2020 Workshop on AI Evaluation, New York, February 7, 2020, https://doi.org/10.48550/arXiv.2102.00813.

[13] ImageNet, https://www.image-net.org (accessed December 12, 2023).

[14] Li Fei-Fei and Ranjay Krishna, "Searching for Computer Vision North Stars," *Dædalus* 151 (2) (Spring 2022): 85–99, https://www.amacad.org/publication/searching-computer-vision-north-stars.

[15] Raji and Fried, "About Face."

[16] Vinay Uday Prabhu and Abeba Birhane, "Large Image Datasets: A Pyrrhic Win for Computer Vision?" paper presented at the Institute of Electrical and Electronics Engineers/Computer Vision Foundation Conference on Applications of Computer Vision, 2021, https://openaccess.thecvf.com/content/WACV2021/papers/Birhane_Large_Image_Datasets_A_Pyrrhic_Win_for_Computer_Vision_WACV_2021_paper.pdf.

[17] Xiang, "Being 'Seen' vs. 'Mis-Seen.'"

[18] Michele Merler, Nalini Ratha, Rogerio Feris, and John R. Smith, "Diversity in Faces," arXiv, April 10, 2019, https://arxiv.org/pdf/1901.10436.pdf.

[19] Taylor Hatmaker, "Lawsuits Allege Microsoft, Amazon and Google Violated Illinois Facial Recognition Privacy Law," Tech Crunch, July 15, 2020, https://techcrunch.com/2020/07/15/facial-recognition-lawsuit-vance-janecyk-bipa.

[20] Nicolas Rivero, "The Influential Project that Sparked the End of IBM's Facial Recognition Program," *Quartz*, June 10, 2020, https://qz.com/1866848/why-ibm-abandoned-its-facial-recognition-program.

[21] Kaiyu Yang, Jacqueline Yau, Li Fei-Fei, et al., "A Study of Face Obfuscation in ImageNet," arXiv, March 10, 2021, last revised June 9, 2022, https://doi.org/10.48550/arXiv.2103.06191; Common Objects in Context, https://cocodataset.org (accessed December 12, 2023); and Julienne LaChance, William Thong, Shruti Nagpal, and Alice Xiang, "A Case Study in Fairness Evaluation: Current Limitations and Challenges for Human Pose Estimation," paper presented at the Association for the Advancement of Artificial Intelligence 2023 Workshop on Representation Learning for Responsible Human-centric AI (R2HCAI), Washington, D.C., February 13, 2023, https://r2hcai.github.io/AAAI-23/files/CameraReadys/21.pdf.

[22] Abhishek Mandal, Susan Leavy, and Suzanne Little, "Dataset Diversity: Measuring and Mitigating Geographical Bias in Image Search and Retrieval," *Proceedings of the First International Workshop on Trustworthy AI for Multimedia Computing* (Trustworthy AI '21): 19–25, https://doi.org/10.1145/3475731.3484956; and Naggita, LaChance, and Xiang, "Flickr Africa."

[23] Angelina Wang, Alexander Liu, Ryan Zhang, et al., "REVISE: A Tool for Measuring & Mitigating Bias in Visual Datasets," paper presented at the European Conference on Computer Vision, virtual, August 23, 2020, https://arxiv.org/pdf/2004.07999.pdf.

[24] Jieyu Zhao et al., "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-Level Constraints," paper presented at the Conference on Empirical Methods in Natural Language Processing, Copenhagen, September 7, 2017, https://arxiv.org/abs/1707.09457; and Dora Zhao, Jerone Andrews, and Alice Xiang, "Men Also Do Laundry: Multi-Attribute Bias Amplification," *Proceedings of Machine Learning Research* 202 (2023): 42000–42017, https://arxiv.org/abs/2210.11924.

[25] Tessa E. S. Charlesworth, Victor Yang, Thomas C. Mann, et al., "Gender Stereotypes in Natural Language: Word Embeddings Show Robust Consistency Across Child and Adult Language Corpora of More Than 65 Million Words," *Psychological Science* (2021): 1–23.

[26] *In re: Facebook Biometric Information Privacy Litigation*, 185 F. Supp. 3d 1155 (N.D. Cal. 2016), Order Re Final Approval, Attorneys' Fees and Costs, and Incentive Awards; Signed by Judge James Donato on February 26, 2021, https://www.govinfo.gov/app/details/USCOURTS-cand-3_15-cv-03747/USCOURTS-cand-3_15-cv-03747-16.

[27] Richard Van Noorden, "The Ethical Questions that Haunt Facial Recognition Research," *Nature* 587 (2020): 354–358, https://doi.org/10.1038/d41586-020-03187-3.

[28] Prabhu and Birhane, "Large Datasets."

[29] Richard D. Taylor, "'Data Localization': The Internet in the Balance," *Telecommunications Policy* 44 (8) (2020), https://doi.org/10.1016/j.telpol.2020.102003.

[30] James Griffiths, "Acronyms and Abbreviations," in *The Great Firewall of China: How to Build and Control an Alternative Version of the Internet* (London: Zed Books Ltd, 2019), xi–xii.

[31] Daniel E. Ho, Emily Black, Maneesh Agrwawala, and Fei-Fei Li, "Domain Shift and Emerging Questions in Facial Recognition Technology," policy brief, Stanford Uni-

versity Human-Centered Artificial Intelligence, https://hai.stanford.edu/sites/default/files/2020-11/HAI_FRT_WhitePaper_PolicyBrief_Nov2020.pdf.

[32] Sidney Fussell, "How an Attempt at Correcting Bias in Tech Goes Wrong," *The Atlantic*, October 29, 2019, https://www.theatlantic.com/technology/archive/2019/10/google-allegedly-used-homeless-train-pixel-phone/599668; and Billy Perrigo, "Inside Facebook's African Sweatshop," *Time*, February 14, 2022, https://time.com/6147458/facebook-africa-content-moderation-employee-treatment.

[33] Xiang, "Being 'Seen' vs. 'Mis-Seen.'"

[34] Ibid.

[35] Marie des Neiges Leonard, "Census and Racial Categorization in France: Invisible Categories and Color-Blind Politics," *Humanity & Society* 38 (1) (2014): 67–88; and Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker, "How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis," *Proceedings of the ACM on Human-Computer Interaction* 4 (CSCW1) (2020): 1–35, https://doi.org/10.1145/3392866.

[36] Jessica L. Roberts, "Protecting Privacy to Prevent Discrimination," *William and Mary Law Review* 56 (6) (2015): 2097–2174, https://scholarship.law.wm.edu/wmlr/vol56/iss6/4.

[37] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," paper presented at the Eighth Conference on Innovations in Theoretical Computer Science, Cambridge, Mass., January 14, 2016, https://arxiv.org/abs/1609.05807.

[38] Sam Corbett-Davies, Emma Pierson, Avi Feller, et al., "Algorithmic Decision Making and the Cost of Fairness," paper presented at the 23rd ACM SIGKDD International Conference on Knowledge, Discovery & Data Mining, Halifax, August 13–17, 2017; Sam Corbett-Davies, Johann D. Gaebler, Hamed Nilforoshan, et al., "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning," *Journal of Machine Learning Research* 24 (2023), https://www.jmlr.org/papers/v24/22-1511.html; Alexandra Chouldechova, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," *Big Data* 5 (2) (2017): 153–163; and James R. Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan, "An Intersectional Definition of Fairness," paper presented at the 36th Institute of Electrical and Electronics Engineers International Conference on Data Engineering, Dallas, Tex., April 21, 2020.

[39] Alice Xiang, "Reconciling Legal and Technical Approaches to Algorithmic Bias," *Tennessee Law Review* 88 (2021): 649–724; Daniel E. Ho and Alice Xiang, "Affirmative Algorithms: The Legal Grounds for Fairness as Awareness," *University of Chicago Law Review Online* (2020), https://lawreviewblog.uchicago.edu/2020/10/30/aa-ho-xiang; Jason R. Bent, "Is Algorithmic Affirmative Action Legal?" *The Georgetown Law Journal* 108 (2020): 803–853; and Zach Harned and Hanna Wallach, "Stretching Human Laws to Apply to Machines: The Dangers of a 'Colorblind' Computer," *Florida State University Law Review* 47 (617) (2020).

[40] Jack M. Balkin and Reva B. Siegel, "The American Civil Rights Tradition: Anticlassification or Antisubordination?" *University of Miami Law Review* 9 (10) (2003).

[41] Pamela Kirkland, "For Howard Grads, LBJ's 'To Fulfill These Rights' Remarks Are Still Relevant Half a Century Later," *The Washington Post*, June 4, 2015, https://www.washingtonpost.com/news/post-nation/wp/2015/06/04/for-howard-grads-lbjs-to-fulfill-these-rights-remarks-are-still-relevant-half-a-century-later.

[42] Xiang, "Reconciling Legal and Technical Approaches to Algorithmic Bias."

[43] *Regents of the University of California v. Bakke*, 438 U.S. 265 (1978); *Grutter v. Bollinger*, 539 U.S. 306 (2003); and *Gratz v. Bollinger*, 539 U.S. 244 (2003).

[44] Xiang, "Reconciling Legal and Technical Approaches to Algorithmic Bias."

[45] *Students for Fair Admissions v. Harvard*, 600 U.S. 181 (2023).

[46] Ibid., 22–24, 30–32.

[47] Ibid., 27.

[48] Ibid., 39–40.

[49] Xiang, "Reconciling Legal and Technical Approaches to Algorithmic Bias."

[50] Ian F. Haney López, "'A Nation of Minorities': Race, Ethnicity and Reactionary Colorblindness," *Stanford Law Review* 59 (4) (2007): 985–1063, http://www.stanfordlawreview.org/wp-content/uploads/sites/3/2010/04/lopez.pdf.

[51] Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," *California Law Review* 104 (3) (2016): 671–732; and Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, et al., "Gender Bias in Word Embeddings: A Comprehensive Analysis of Frequency, Syntax, and Semantics," paper presented at the Association for the Advancement of Artificial Intelligence/Association for Computing Machinery Conference on AI, Ethics, and Society, Oxford, August 1, 2022.

[52] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias," *ProPublica*, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[53] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang, "What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness," *Proceedings of the 2021 Association for Computing Machinery Conference on Fairness, Accountability, and Transparency*, 249–260, https://doi.org/10.1145/3442188.3445888.

[54] George E. P. Box, "Science and Statistics," *Journal of the American Statistical Association* 71 (356) (1976): 791–799.

[55] A. Feder Cooper and Ellen Abrams, "Emergent Unfairness in Algorithmic Fairness-Accuracy Trade-Off Research," *Proceedings of the 2021 Association for the Advancement of Artificial Intelligence/Association for Computing Machinery Conference on AI, Ethics, and Society*, 46–64, https://doi.org/10.1145/3461702.3462519.

[56] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science* 366 (6464) (2019): 447–453, https://doi.org/10.1126/science.aax2342.

[57] Ho and Xiang, "Affirmative Algorithms."

[58] Anastasia Kozyreva, Stefan M. Herzog, Stephan Lewandosky, et al., "Resolving Content Moderation Dilemmas between Free Speech and Harmful Misinformation," *Proceedings of the National Academy of Sciences* 120 (7) (2023): e2210666120, https://www.pnas.org/doi/abs/10.1073/pnas.2210666120.

[59] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite, "Stable Bias: Analyzing Societal Representations in Diffusion Models," arXiv, March 20, 2023, last revised November 9, 2023, https://arxiv.org/abs/2303.11408.

[60] Xiang, "Being 'Seen' vs. 'Mis-Seen.'"

[61] Salomé Viljoen, "A Relational Theory of Data Governance," *Yale Law Journal* 131 (2) (2021): 573–654, https://www.yalelawjournal.org/pdf/131.2_Viljoen_1n12myx5.pdf.

[62] Antoaneta Roussi, "Resisting the Rise of Facial Recognition," *Nature* 587 (2020): 350–353, https://www.nature.com/articles/d41586-020-03188-2.

[63] See, for example, Kashmir Hill, "Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match," *The New York Times*, December 29, 2020, https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html; Kashmir Hill, "Wrongfully Accused by an Algorithm," *The New York Times*, June 24, 2020, https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html; and Elisha Anderson, "Controversial Detroit Facial Recognition Got Him Arrested for a Crime He Didn't Commit," *Detroit Free Press*, July 10, 2020, https://www.freep.com/story/news/local/michigan/detroit/2020/07/10/facial-recognition-detroit-michael-oliver-robert-williams/5392166002.

Deprogramming Implicit Bias: The Case for Public Interest Technology

Darren Walker

New technologies have fundamentally transformed the systems that govern modern life, from criminal justice to health care, housing, and beyond. Algorithmic advancements promise greater efficiency and purported objectivity, but they risk perpetuating dangerous biases. In response, the field of public interest technology has emerged to offer an interdisciplinary, human-centered, and equity-focused approach to technological innovation. This essay argues for the widespread adoption of public interest technology principles, including thinking critically about how and when technological solutions are deployed, adopting rigorous training to educate technologists on ethical and social context, and prioritizing the knowledge and experiences of communities facing the disproportionate harms or uneven benefits of technology. Tools being designed and deployed today will shape our collective future, and collaboration between philanthropy, government, storytellers, activists, and private-sector technologists is essential in ensuring that these new systems are as just as they are innovative.

Three years ago, Robert Julian-Borchak Williams, a Detroit office worker, received a call from the Detroit Police Department. He assumed it was a prank, but when he pulled into his driveway, police officers were waiting in his front yard. They handcuffed Robert in front of his wife and daughters, and refused to answer his family's panicked questions. Williams spent the night in a crowded jail cell. The next afternoon, the day before his forty-second birthday, the police brought him to an interrogation room. Stone-faced detectives showed him photographs of a robbery suspect. "Is this you?" they demanded. Williams held the photograph next to his face. The image clearly displayed a different man. The reason for Williams's unjustified arrest was not a witness statement or a botched DNA match. Instead, Williams had been falsely identified by law enforcement officers who used a faulty facial recognition algorithm to ensnare the wrong man in the criminal legal system.[1] While Robert Williams's story is alarming, it is not an anomaly. Since the Detroit Police Department began using facial recognition, at least two other Black men in the same city have been falsely arrested, destroying their job prospects and fracturing a marriage.[2] One of these men even considered

https://doi.org/10.1162/daed_a_02059

accepting a plea deal for a crime he did not commit. In fact, Detroit's facial algorithm misidentifies suspects more than 90 percent of the time.[3] Yet it is still used widely across the department, nearly exclusively against Black people. In Detroit, as elsewhere across the country, technology replicates, reinforces, and indeed masks human bias on a scale we have never encountered before, a scale only accessible in the language of machines. Algorithms, artificial intelligence, and technology pervade our criminal legal system, often with little oversight. Judges use risk-assessment technology to determine parole and probation terms.[4] Compared with white defendants, some of these tools are 77 percent more likely to predict that Black defendants will commit a violent offense.[5]

These harmful algorithms extend beyond the criminal legal system, to the services that determine health and safety. An algorithm used to manage health care for two hundred million people in the United States was found to refer disproportionately few Black people to programs providing personalized care, even though Black patients were often substantially more ill than their white counterparts.[6] Meanwhile, landlords across the country increasingly rely on artificial intelligence to screen applicants, including with algorithms that can penalize applicants for criminal accusations that are later dropped.[7] Even issues as mundane as the photos we see on our screens are affected by biased technology. In one widely cited example, a Google Photos algorithm falsely identified Black people as gorillas.[8] Technologies that once seemed confined to science-fiction novels are now embedded in our democracy, and with them, a host of algorithmic biases at a colossal and concerning scale. These examples, among many others, indicate a recursive problem. Our algorithms are embedded with the biases of the humans who create them; and with each additional algorithm built atop an unjust foundation, the initial bias recurs, repeats, and worsens, to devastating effect.

When privatized, without oversight and careful regulation, this self-sustaining cycle of algorithmic bias will continue unabated, not only exacerbating existing inequality but creating new inequalities altogether. As Latanya Sweeney, head of Harvard's Public Interest Tech Lab and former chief technology officer of the U.S. Federal Trade Commission, rightly noted, "Once a design or business practice works, it gets replicated just as it is. The design of the technology really does dictate the rules that we have to live by."[9] Those of us invested in a more just and equitable future face an urgent question: How do we address this mounting crisis of algorithmic injustice?

Some argue that the project of reforming technology is best left in the hands of programmers and specialists: the technical experts who designed these systems. As technology advances, this logic contends, its consequences will reveal themselves, and then be corrected by the forward march of new technology. Certainly, these groups have crucial expertise and insight needed to understand the algo-

rithms that define our lives. But the growth-at-any-cost mindset that pervades the tech industry often overlooks the realities of race, gender, and disability inequities, and risks repeating a vicious cycle ad infinitum.[10]

On the other end of the spectrum, a coalition of industry leaders and technologists recently signed a letter calling for an AI development moratorium.[11] This short-term solution would do little to address the structural issues that shape the development of artificial intelligence. For instance, while it might tackle discrete safety concerns, it is unlikely to fundamentally shift the training that computer scientists and engineers receive to grapple with technology's unintended consequences for marginalized groups. A tech-imposed temporary stoppage also problematically implies that the industry is self-governed, which is simply not true. Existing federal regulatory schemes, from product liability statutes to civil rights protections, already apply to artificial intelligence.[12] The answer is not to ask for a proverbial time out, but rather to bring in the referees: the advocates and regulators who carry the capacity and technical expertise to enforce laws and correct violations at scale. Moving forward, we should address this recursive problem the way we would any other: by breaking it down into a series of smaller subproblems and solving them one at a time. We might start by investing in the excellence of a new generation of talented technologists with the technical expertise, interdisciplinary training, and lived experience to deploy strategies that end algorithmic bias, once and for all.

The good news is academics, advocates, and technologists have been engaged in this work for years, building the new field of public interest technology together. This interdisciplinary approach calls for technology to be designed, deployed, and regulated in a responsible and equitable manner.[13] It goes beyond designing technology for good, asking and answering: "Good for *whom*?" Public interest technologists center *people,* not innovation for its own sake. They focus on those most affected by new innovations: the historically marginalized groups who have experienced the most harms or the uneven benefits of technology. At the same time, public interest technologists understand that technology is not, and never has been, neutral. The dangers of technology, they argue, cannot be resolved with one product or program. Instead, these technologists evaluate and address potential inequalities at every stage of innovation, from design and development to the real-world impact in the hands of users. The field includes leading technical experts, researchers, and scientists. And it invites those outside of technology – storytellers, activists, artists, and academics – to offer their crucial expertise and hold designers and decision-makers accountable. As celebrated filmmaker Ava DuVernay noted about the artist's role in addressing these harms: "The idea that the story that technology is telling about us could possibly not be our true story, makes it just as important as any crime thriller I might be

covering."[14] Simply put, public interest technology is a multisector effort. It calls everyone to consider how we use, encourage, and adopt technology in our lives, our fields, and our broader institutions.

From academics to funders to private-sector innovators, we can all benefit from taking a public interest technology approach to our work. First, we can and must question the gospel of tech solutionism.[15] Instead of assuming new technology will inevitably correct a social ill, we must think more critically about how and when technology is deployed. Being more intentional about the technology we adopt can move us from *reacting* to unforeseen consequences to *preventing* these negative effects. For example, the Algorithmic Justice League, an organization devoted to "unmasking AI harm," and other advocates recently prevented the Internal Revenue Service from implementing a controversial plan forcing taxpayers to use facial-recognition software to log in to their IRS accounts.[16] The change would have exposed millions to privately owned software with limited oversight.

Second, we must also embed rigorous public interest technology training in computer science, engineering, and data science curriculums. Such training will ensure that talented technologists graduate with both technical expertise and an extensive understanding of the social context in which technology is deployed. These efforts may also include funding or pursuing research and projects that interrogate how technology furthers systemic bias.[17] Such revelations have come from resource hubs like those at Harvard's Public Interest Tech Lab.[18] Researchers and scientists at the lab have unmasked biased Facebook advertising algorithms that targeted Black users and exposed the proliferation of deepfake comments in U.S. public comment sites.[19] And educational institutions nationwide are building the next generation of public interest technologists – together. The Public Interest Technology University Network unites sixty-three universities in connecting public interest students and faculty with resources and institutional support.[20]

Of course, any attempt to correct technology's ills will fall short if we do not center the knowledge and experiences of the marginalized people most vulnerable to its inherent risks. So, technologists can and must partner with marginalized communities to repair the damage caused by bias and prevent it from the outset. For instance, after studies revealed that non-white Airbnb hosts were earning less money than their white counterparts, Airbnb partnered with civil rights organizations to create Project Lighthouse, an initiative to reduce discrimination for hosts and travelers on the platform.[21] These efforts drew on the experiences of Black hosts and guests, who shared their struggles with securing housing under the hashtag #AirbnbWhileBlack.[22]

Finally, public interest technologists themselves can and must draw on their own intersectional experiences, with support from funders and academic institutions alike. At the Ford Foundation, our commitment to public interest technol-

ogy arose out of a strategy to promote internet rights and digital justice. Through our Technology and Society program, Ford has committed more than $100 million to fostering the field of public interest technology since 2016 – all to build an ecosystem that will lead to a more just technological future for all. Many researchers affiliated with the program have personally experienced the harms of biased algorithms or inaccessible technology. They bridge specialized expertise with a rich personal background, advocating for structural and long-term solutions like an AI Bill of Rights, which would ensure that a shared set of norms and values shape technology to better serve the public good.[23]

Technology's ever-changing landscape presents a daunting challenge. Nevertheless, I am hopeful for a future in which technology empowers us to serve the public good, because I know we've solved these problems before. Indeed, the ideological ancestor of the public interest technology field exists. It is called public interest law.

Six decades ago, during the early 1960s, there was no such thing as public interest law. Law schools focused on academic and corporate issues to the detriment of addressing social inequities. Legal aid groups struggled to survive. But the Ford Foundation set out to change that and to train a new generation of lawyers who would work in the best interest of the public to provide legal representation to low-income and marginalized groups, engage in advocacy more broadly, and expand rights throughout society. By the time I graduated from law school in the mid-1980s, the once-nascent field was flourishing. Today, public interest law is so prominent that many take it for granted. Low-income tenants who have been evicted can join a class-action lawsuit, free of charge. Young people fleeing discriminatory anti-LGBTQ+ legislation can access entire organizations dedicated to supporting their legal rights. The field is far from perfect but it's a prescient reminder that time, investment, and collaboration can turn a sore lack into a surplus. Those who have long driven the field of public interest law – people of color, people with disabilities, low-income people, and LGBTQ+ people – are best equipped to fight a barrage of implicit bias–based challenges. If we support them, we can build a parallel public interest field anew.

The technology that determines our housing, health, and safety cannot and must not be the protected intellectual property of a few. It is a public good for the many. And people from every sector can contribute to a more just vision of tech by extending support and funding for crucial research, welcoming public interest technologists to nontechnical fields, and advancing solutions that reject the philosophy of "move fast and break things" by instead calling us all to fix what is broken.[24] By embarking on this mission to center people in the technology that is supposed to help us, we move toward justice for the millions of people who face algorithmic bias in their everyday lives, including Robert Williams, who is still reckoning with the consequences of his false arrest. It has been three years since Williams was wrongly

handcuffed on his front lawn, but his seven-year-old daughter still cries when she sees his arrest footage.[25] And still the recursive loop circles.

On November 25, 2022, Randal Reid, a Black man, was driving in Georgia to a late Thanksgiving celebration with his mother. Police pulled him over, announcing there was a warrant for his arrest for a theft that had occurred in Louisiana.[26] Reid pleaded that he had never spent a day in Louisiana. Yet he was booked and spent six days in jail based on an incorrect facial recognition match claiming he was a man forty pounds heavier and without a mole on his face. Let us learn with humility from the shattering experiences endured by too many families and break this recursive loop before it's too late.

ABOUT THE AUTHOR

Darren Walker, a Fellow of the American Academy since 2015, is President of the Ford Foundation. He previously served as Vice President of the Rockefeller Foundation, where he oversaw the Rebuild New Orleans initiative after Hurricane Katrina. He is the author of *From Generosity to Justice: A New Gospel of Wealth* (2023) and has published in journals such as *JAMA* and *Foreign Affairs*.

ENDNOTES

[1] Kashmir Hill, "Wrongfully Accused by an Algorithm," *The New York Times*, June 24, 2020, https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html.

[2] Khari Johnson, "How Wrongful Arrests Based on AI Derailed 3 Men's Lives," *Wired*, March 7, 2022, https://www.wired.com/story/wrongful-arrests-ai-derailed-3-mens-lives.

[3] Jason Koebler, "Detroit Police Chief: Facial Recognition Software Misidentifies 96% of the Time," *Vice*, June 29, 2020, https://www.vice.com/en/article/dyzykz/detroit-police-chief-facial-recognition-software-misidentifies-96-of-the-time.

[4] Michael Brenner, Jeannie Suk Gersen, Michael Haley, et al., "Constitutional Dimensions of Predictive Algorithms in Criminal Justice," *Harvard Civil Rights-Civil Liberties Law Review* 55 (1) (2020): 267–310.

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias," *ProPublica*, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; and Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin, *ProPublica*, May 23, 2016, "How We Analyzed the COMPAS Recidivism Algorithm," https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[6] Heidi Ledford, "Millions of Black People Affected by Racial Bias in Health-Care Algorithms," *Nature* 574 (7780) (2019): 608–609, https://doi.org/10.1038/d41586-019-03228-6.

[7] Valerie Schneider, "Locked Out by Big Data: How Big Data, Algorithms, and Machine Learning May Undermine Housing Justice," *Columbia Human Rights Law Review* 52 (1) (2020): 251–305.

[8] Jenna Burrell, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms," *Big Data & Society* 3 (1) (2016): 1–12, http://doi.org/10.2139/ssrn.2660674.

[9] Dave Gershgorn, "A Harvard Professor Thinks that Tech's True Power Comes from Design," *Quartz*, February 24, 2018, https://qz.com/1214645/latanya-sweeney-explains-why-tech-companies-are-so-powerful.

[10] Greta Byrum and Ruha Benjamin, "Disrupting the Gospel of Tech Solutionism to Build Tech Justice," *Stanford Social Innovation Review*, June 16, 2022, https://ssir.org/articles/entry/disrupting_the_gospel_of_tech_solutionism_to_build_tech_justice.

[11] Cade Metz and Gregory Schmidt, "Elon Musk and Others Call for Pause on A.I., Citing 'Profound Risks to Society,'" *The New York Times*, March 29, 2023, https://www.nytimes.com/2023/03/29/technology/ai-artificial-intelligence-musk-risks.html.

[12] See Charlotte A. Burros, Rohit Chopra, Kristen Clarke, and Lina M. Khan, "Joint Statement on Enforcement Efforts Against Discrimination and Bias in Automated Systems," Federal Trade Commission, April 25, 2023, https://www.ftc.gov/legal-library/browse/cases-proceedings/public-statements/joint-statement-enforcement-efforts-against-discrimination-bias-automated-systems.

[13] Katharine Lusk, "Public Interest Technology University Network: Understanding the State of the Field," Boston University Initiative on Cities, May 31, 2022, https://open.bu.edu/handle/2144/44469.

[14] See Public Interest Technology University Network, "Using Tech for Good: A New Generation of Civic-Minded Technologies," filmed October 7, 2019, at Georgetown University, Washington, D.C., video, [38:45], https://www.youtube.com/watch?v=cP1IMsFQMDQ&t=2s.

[15] Byrum and Benjamin, "Disrupting the Gospel of Tech Solutionism to Build Tech Justice."

[16] Rachel Metz, "Activists Pushed the IRS to Drop Facial Recognition. They Won, but They're Not Done Yet," CNN Business, March 7, 2022, https://amp.cnn.com/cnn/2022/03/07/tech/facial-recognition-activists-irs/index.html; and Joy Buolamwini, "The IRS Should Stop Using Facial Recognition," *The Atlantic*, January 27, 2022, https://www.theatlantic.com/ideas/archive/2022/01/irs-should-stop-using-facial-recognition/621386.

[17] Lusk, "Public Interest Technology University Network."

[18] See Harvard University, "Public Interest Tech Lab," https://techlab.org.

[19] Jinyan Zang, "How Facebook's Advertising Algorithms Can Discriminate By Race and Ethnicity," *Technology Science*, October 19, 2021, https://techscience.org/a/2021101901; and Max Weiss, "Deepfake Bot Submissions to Federal Public Comment Websites Cannot Be Distinguished from Human Submissions," *Technology Science*, December 17, 2019, https://techscience.org/a/2019121801/#Citation.

[20] See New America, "The Public Interest Technology University Network," https://pitcases.org.

[21] Benjamin Edelman, Michael Luca, and Dan Svirsky, "Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment," *American Economic Journal: Applied Economics* 9 (2) (2017): 1–22, https://doi.org/10.1257/app.20160213; and Airbnb, "A New Way

We're Fighting Discrimination on Airbnb," June 15, 2020, https://www.airbnb.com/resources/hosting-homes/a/a-new-way-were-fighting-discrimination-on-airbnb-201.

[22] Maggie Penman, Shankar Vedantam, and Max Nesterak, "#AirbnbWhileBlack: How Hidden Bias Shapes the Sharing Economy," NPR, April 26, 2016, https://www.npr.org/2016/04/26/475623339/-airbnbwhileblack-how-hidden-bias-shapes-the-sharing-economy.

[23] The White House, "Blueprint for an AI Bill of Rights," October 2022, https://www.whitehouse.gov/ostp/ai-bill-of-rights.

[24] Hemant Taneja, "The Era of 'Move Fast and Break Things' Is Over," *Harvard Business Review*, January 22, 2019, https://hbr.org/2019/01/the-era-of-move-fast-and-break-things-is-over.

[25] Johnson, "How Wrongful Arrests Based on AI Derailed 3 Men's Lives."

[26] Kashmir Hill and Ryan Mac, "'Thousands of Dollars for Something I Didn't Do,'" *The New York Times*, March 31, 2021, https://www.nytimes.com/2023/03/31/technology/facial-recognition-false-arrests.html.

Beyond Implicit Bias

Thomas D. Albright, William A. Darity Jr., Diana Dunn, Rayid Ghani, Deena Hayes-Greene, Tanya Katerí Hernández & Sheryl Heron

In their introduction to this edition of *Dædalus*, Goodwin Liu and Camara Phyllis Jones write that "it is unlikely that implicit bias can be effectively addressed by cognitive interventions alone, without broader institutional, legal, and structural reforms." They note that the genesis for the volume was a March 2021 workshop on the science of implicit bias convened by the Committee on Science, Technology, and Law of the National Academies of Sciences, Engineering, and Medicine.[1] That workshop provided an opportunity to demonstrate that implicit bias is a common form of cognitive processing that develops in response to social, cultural, and institutional conditions. As demonstrated by the workshop and the essays in this volume, an understanding of implicit bias in a neurological, mechanistic, and phenomenological manner strengthens our ability to develop policies to diffuse and mitigate the problems that arise from implicit bias.

At the end of the 2021 event, members of the interdisciplinary workshop planning committee gave their perspectives on the important messages that they would take away from the workshop. For the conclusion of this volume of *Dædalus*, we members of the planning committee were asked to expand on what we said three years ago. This is our response.

Thomas D. Albright

Broadly considered, implicit bias is a cognitive response to uncertainty, in which other pieces of information are unconsciously recruited to fill in the blanks of experience. This inferential process is probabilistic and sometimes yields catastrophic outcomes. This is particularly true in a human social context, in which uncertainty is pervasive and other pieces of information include tribal allegiances and social structures that yield disparate treatment as a function of race. Unconscious incorporation of this information allows it to be manifested as implicit racial bias.

https://doi.org/10.1162/daed_a_02060

The essays in this issue of *Dædalus* catalog the incidence of implicit biases and their devastating effects on individual opportunities and social cohesion. They also explore the societal forces and mechanisms responsible for the development and perseveration of biases. This evidence-based understanding sets the stage for the most important question: how do we stop this from happening?

There are three information-processing strategies that hold promise: 1) reduce uncertainty, 2) change the priors, and 3) compensate. As seen from the essays in this volume, there has been significant growth of science that tests the effectiveness of these strategies.

Reduce uncertainty. Unconscious biases flourish where there is paucity of information. In the case of implicit racial bias, this comes from long-standing forms of cultural and geographic segregation. Evidence suggests that uncertainty can be reduced by engineering meaningful interactions between people from different racial groups, such that information about the "unknown other" is acquired broadly over time and different contexts.

Change the priors. Implicit bias is a form of statistical inference based on observed events and associations. Explicit racism in American society provides a model from which generations of children have acquired a distorted sense of the character of people of different races. As long as that model exists, our priors contain incorrect information, yielding unconscious bias. Hope lies in the fact that acquiring new associations predictably alters unconscious inference, manifested as changes in perception, decision, and action.

Compensate. Efforts to reduce uncertainty and change flawed priors are long-term solutions. Along the way, one valuable strategy is to recognize the biases we hold and overpower them. Because this compensation requires a rational conscious consideration of the potential for error under normal conditions of unconscious bias, the simplest and perhaps most immediately effective strategy is to pause and think before a decision to act. Implicit bias training commonly focuses on this moment, in which qualified decision-making can prevent the harmful biases we have acquired, and hope to suppress, from having real impact on the world in front of us.

William A. ("Sandy") Darity Jr.

The related concepts of unconscious bias and implicit bias have potential value in analyzing personal interactions fraught with prejudice. The two concepts enable individual bigoted acts predicated upon stereotypical beliefs to be viewed as devoid of intent or malice. Both concepts can improve our understanding of interpersonal racism.

In contrast, unconscious bias and implicit bias are far less useful in understanding structural racism, those social practices and policies that produce and

sustain racial inequality. Those practices and policies have been constructed and maintained in both conscious and explicit fashion by their designers.

For example, the incorporation of decentralized authority of the administration of the GI Bill enacted after World War II was an intentional measure to benefit white veterans at the expense of Black veterans. In the late nineteenth century, the failure of the federal government to fulfill its promise of forty-acre land grants to the formerly enslaved as restitution for their years of bondage, while mobilizing the Homestead Act of 1862 to provide 1.5 million white families with 160-acre land grants in the Western territories, was deliberate and purposeful. In the 1950s through the 1970s, the grossly disproportionate placement of freeways under the federal highway system in the heart of predominantly Black neighborhoods and business districts was calculated and willful.

Of course, there have been policies adopted for purposes other than preserving racial hierarchy that have had an inordinate adverse effect on Black community well-being. However, those effects could have been anticipated and mitigated had a careful racial impact audit been performed in advance of their introduction. The failure to conduct such an audit has been the product of a conscious decision by policymakers.

Diana Dunn

In volume one of *Undoing Racism: A Philosophy of International Social Change*, psychologists Ronald Chisholm and Michael Washington define race as "a specious classification of human beings created by Europeans (whites) which assigns human worth and social status using 'white' as the model of humanity and the height of human achievement for the purpose of establishing and maintaining privilege and power."[2]

When white people struggle, we change structures, but when Black people struggle, we give them programs that function within current structures. Only by empowering those most impacted by racism can we create movements that lead to meaningful change.

Rayid Ghani

As we discuss implicit bias in humans, we also need to consider, understand, and deal with that implicit bias propagating in society through the design and use of AI systems. As AI systems augment various existing human processes in society through widespread use by governments and corporations, developing approaches to better understand, detect, and deal with them is a critical need for society.

AI systems have the potential to help improve outcomes and result in a better and more equitable society across a broad range of areas including social welfare, health, education, and criminal justice. At the same time, any AI (or otherwise developed) system that affects people's lives must be explicitly built to increase equity and focus on tackling the implicit (and sometimes explicit) biases underlying the design choices made in the development of that system. It is important to recognize that AI can have a positive social impact, but we need to make sure that we put guidelines, resources, and trainings in place to maximize the chances of the positive impact while protecting people who have historically been negatively impacted by implicit bias in society, and will likely continue to be affected negatively by the new AI systems. This requires government agencies, businesses, nonprofits, and community groups to come together and collaborate around this goal, and for policymakers to act and provide guidelines and/or regulations for both the public- and private-sector organizations using AI-assisted decision-making processes to ensure that these systems are built in a transparent and accountable manner and result in fair and equitable outcomes for society.

When designed appropriately and deliberately, AI can be a useful tool to assist us in both detecting implicit biases when they're present in a human process, as well as in designing new collaborative systems that help reduce the impact of these biases.

Deena Hayes-Greene

Cultural and racial biases are often thought to be an indication of racial prejudice or bigotry – and they can be. However, they are as likely to be the result of the associations our brains make, many of which we are unaware. These associations (for example, shark-danger, fire-hot) are made to save our lives. They become racial when our focus is not on fire or sharks but on people. People whom we might not have met but who are associated with danger or negative labels ascribed by a society, a nation, established in more racially exclusive times. These associations have been nourished for centuries by a culture designed to advantage white people. And the structures such a culture builds and sustains successfully separate and harm poor and working-class people of all races. Drilled into our heads are long-established, time-honored associations about who is valuable, worthy, and deserving. Race then remains the ever-ready tool to prompt well-rooted but implicit associations. Far from the prejudices of individuals, the constant repetition of a hierarchy of human worth, when commonly held, directs the very construction of today's world and creates disparity.

Without interruption, such associations remake themselves with each generation because racism is a toxin in our nation's groundwater. In other words, if you

come upon a lake and see one fish belly up, dead, examining the fish and seeing what caused its death would make sense. If, however, you came upon the lake and half the fish were belly up, what might it be time to do? It is, of course, time to examine the lake. Fish represent the individuals for whom we care, and lakes represent institutions whose toxins could well have caused the need for care. This is possible because lakes are not stand-alone bodies of water. They are fed by the groundwater. The groundwater is the unseen water that makes up 95 percent of the fresh water on our planet. When infected by racism, the groundwater carries it to many lakes, causing many problems. Short of that understanding, our decisions, interventions, and even our visions are misguided, as they fail to see the depth of the problem.

Were we to remove the toxic racial structures underpinning our society, no racial associations would be made. Racial stratification would no longer exist, and racial disparity would be a relic of the past. Racism would no longer inform policy, practice, law, or cultural norms, be our associations explicit or implicit. When we understand not only the brain processes enabling racial associations but also why the associations are there to be made, we can face the past and end its legacy.

Tanya Katerí Hernández

When legal scholars and lawyers consider the literature on implicit bias, they do so for pragmatic reasons. Their principal interest is the desire to enhance the antidiscrimination law project of identifying and addressing discrimination.[3] Indeed, it was the enactment of the Civil Rights Act of 1964's ban on employment discrimination that inspired early iterations of social science–informed workplace trainings.[4]

Contemporary diversity trainings have largely turned to focusing on concerns with implicit bias.[5] However, the trainings for the most part have emphasized the relevance of implicit bias for the individual, and not its implications for structural racism.[6] Yet it is structural racism that antidiscrimination law is geared to address in the context of crafting institutional remedies for findings of discrimination.

It is thus quite heartening that the National Academies' implicit bias workshop not only explained implicit bias, but also linked it to systemic problems. Importantly, two-thirds of human resource specialists report that individual-focused trainings have no effect on the careers of people of color or diversity within the ranks of management, and little effect on levels of implicit bias.[7] Concrete interventions focused on systemic and structural issues make the difference between good and bad implicit bias training.[8]

When training is framed as pertaining to systemic problems and then coupled with complementary measures that engage decision-makers in seeking structur-

al interventions for those systemic problems, workplace diversity is markedly increased as a matter of hiring, retention, and promotion.[9] An example that captured national attention provides a useful illustration. On April 18, 2018, Starbucks employees called the Philadelphia police emergency line to request aid. The cause? Two Black men sitting at a table without placing an order as they waited for the third member of their party to arrive for a meeting. The police arrested the men for "trespassing" and escorted them out of Starbucks in handcuffs. No other White patrons sitting at tables received the same treatment. After cell phone footage of the incident caused a public uproar, Starbucks issued a public apology and closed more than eight thousand U.S. stores for an afternoon of racial bias training for one hundred and seventy-five thousand employees. Notably, the training was accompanied by a structural policy change to disrupt the operation of implicit bias. The new policy states that "any customer is welcome to use Starbucks spaces, including our restrooms, cafes and patios, regardless of whether they make a purchase."[10]

Including concerns about systemic racism in the implicit bias training at Starbucks helped create company support for the structural change with the greatest efficacy for containing the harm of implicit bias. As such, this particular Starbucks effort can serve as a model for how consumers of implicit bias training should encourage program facilitators to speak to systemic and structural issues.[11]

Sheryl Heron

In my professional career, I have been both an emergency physician, where I worked to "stop the bleeding" for patients whom I see for traumatic events, and a public health practitioner, with a focus on injury prevention. While we have been trying to address implicit bias issues at the individual level, the problem is in the prevention of racism at the systemic level. As an immigrant from Jamaica and a Black woman, I call our attention to consider the impact of the caste system. We simply cannot stop at implicit bias or racism. We need to consider the role of caste in this country. Pulitzer Prize–winning author Isabel Wilkerson notes that racism is an insufficient term for the systemic oppression of Black people in America and we must consider the caste system that is a part of this country.[12] In the end, we must ensure we show compassion and empathy in the way we treat one another. We need to move beyond implicit bias, and we must be focused and intentional about how we change the narrative.

ABOUT THE AUTHORS

Thomas D. Albright, a Fellow of the American Academy since 2003, is Professor and Director of the Vision Center Laboratory and Conrad T. Prebys Chair in Vision Research at the Salk Institute for Biological Studies. He has written numerous articles, which have appeared in such journals as *Neuron*, *Journal of Neurophysiology*, and *Proceedings of the National Academy of Sciences of the United States of America*.

William A. ("Sandy") Darity Jr. is the Samuel DuBois Cook Professor of Public Policy, African and African American Studies, and Economics and Director of the Samuel DuBois Cook Center on Social Equity at Duke University. He is the author of *From Here to Equality: Reparations for Black Americans in the Twenty-First Century* (with A. Kirsten Mullen, 2022) and *Persistent Disparity: Race and Economic Inequality in the United States Since 1945* (with Samuel L. Meyers, 1998).

Diana Dunn is the Core Trainer and Organizer at People's Institute for Survival and Beyond.

Rayid Ghani is a Distinguished Career Professor in the Machine Learning Department and the Heinz College of Information Systems and Public Policy at Carnegie Mellon University. He has published in such journals as *Journal of Intelligent Information Systems*, *The British Medical Journal*, and *American Journal of Public Health*.

Deena Hayes-Greene is the Cofounder and Managing Director of Racial Equity Institute, LLC.

Tanya Katerí Hernández is the Archibald R. Murray Professor of Law at Fordham University School of Law, and its Associate Director of the Center on Race, Law, and Justice. She is the author of many articles and the books *Racial Innocence: Unmasking Latino Anti-Black Bias and the Struggle for Equality* (2022), *Multiracials and Civil Rights: Mixed-Race Stories of Discrimination* (2018), and *Racial Subordination in Latin America: The Role of the State, Customary Law and the New Civil Rights Response* (2012).

Sheryl Heron is Professor of Emergency Medicine, Vice-Chair of Faculty Equity, Engagement, and Empowerment for Emergency Medicine, and Chief Diversity and Inclusion Officer, as well as Associate Dean for Community Engagement, Equity, and Inclusion at Emory School of Medicine. She is the editor of *Diversity and Inclusion in Quality Patient Care*, first edition (with Anna Walker Jones, Lisa Moreno-Walton, and Marcus L. Martin, 2016) and second edition (Marcus L. Martin, Lisa Moreno-Walton, and Michelle Strickland, 2019).

ENDNOTES

[1] Goodwin Liu and Camara Phyllis Jones, "Introduction: Implicit Bias in the Context of Structural Racism," *Dædalus* 153 (1) (Winter 2024): 12, https://www.amacad.org/publication/introduction-implicit-bias-context-structural-racism.

[2] Ronald Chisholm and Michael Washington, *Undoing Racism: A Philosophy of International Social Change*, Volume 1 (New York: People's Institute Press, 1997), 30–31.

[3] Charles R. Lawrence III, "The Id, the Ego, and Equal Protection: Reckoning with Unconscious Racism," *Stanford Law Review* 39 (1987): 317–388.

[4] Lily Zheng, *Deconstructed DEI: Your No-Nonsense Guide to Doing the Work and Doing It Right* (Oakland: Berrett-Koehler Publishers, 2023), 149.

[5] Jennifer Y. Kim, "I'm Biased and So Are You. What Should Organizations Do? A Review of Organizational Implicit-Bias Training Programs," *Consulting Psychology Journal* 74 (2022): 19–39.

[6] Jesse Singal, *The Quick Fix: Why Fad Psychology Can't Cure Our Social Ills* (New York: Farrar, Straus and Giroux, 2021), 193.

[7] Frank Dobbin and Alexandra Kalev, "Why Doesn't Diversity Training Work? The Challenge for Industry and Academia," *Anthropology Now* 10 (2018): 48–55.

[8] Alexander Kalev, Frank Dobbin, and Erin Kelly, "Best Practices or Best Guesses? Assessing the Efficacy of Corporate Affirmative Action and Diversity Policies," *American Sociological Review* 71 (4) (2006): 589–617.

[9] Tanya Katerí Hernández, "Can CRT Save DEI: Workplace Diversity, Equity & Inclusion in the Shadow of Anti-Affirmative Action," *UCLA Law Review Discourse* 71 (forthcoming 2024).

[10] Starbucks (@Starbucks), "We want our stores to be the third place, a warm and welcoming environment where customers can gather and connect. Any customer is welcome to use Starbucks spaces, including our restrooms, cafes and patios, regardless of whether they make a purchase. https://sbux.co/2IVVAJ8," Twitter, May 29, 2018, https://twitter.com/Starbucks/status/1001584492249731072.

[11] It is important to note that one can appreciate Starbucks's DEI efforts, while at the same time acknowledging the critique of Starbucks's consumers boycotting in opposition to the company's dealings with its employee union and its stance with regards to the crisis in Gaza. Omar Mohammed, "Are McDonald's, Starbucks Boycotts Working?" *Newsweek*, November 17, 2023, https://www.newsweek.com/mcdonalds-starbucks-boycotts-israel-hamas-war-1844933.

[12] Isabel Wilkerson, *Caste: The Origins of Our Discontents* (New York: Random House, 2020).

Inside back cover: (top) **Law enforcement.** A traffic stop. Photograph by iStock.com/Ivan Pantic.

(middle) **Employment.** Attendees at a Veteran Employment and Resource Fair in Long Beach, California. Photograph © 2024 by Eric Thayer/Bloomberg via Getty Images.

(bottom) **Courts.** A jury. Photograph by iStock.com/ImageSource.


POLICE


NOW HIRING!
This list is updated every Friday!
Questions? Call us!
562-570-0205
Civil Engineer
Civil Engineering Associate
Combination Building Inspector
Equipment Mechanic I-II
Gardener I-II
Hazardous Material Specialist I-II
Mechanical Engineer
Mechanical Engineering Analyst I-IV
Office Systems Analyst I-IV
Police Officer - Academy Graduate
Police Officer - Lateral
Police Officer - Academy Enrolled
Public Health Nurse I-III
Public Safety Dispatcher
School Guard
Senior Animal Control Officer
Senior Combination Building Inspector
Senior Electrical Inspector
Social Services Officer I-IV
Survey Technician
Tree Trimmer I-II
Water Treatment Operator I-IV

on the horizon:

Advances & Challenges in International Higher Education
edited by Wendy Fischman, Howard Gardner & William C. Kirby

with Wen-hsin Yeh, Isak Frumin, Daria Platonova, Gökhan Depo, Jamshed Bharucha, Tarun Khanna, Takehiko Kariya, Mariët Westermann, Haiyan Gao, Yijun Gu, Marwan M. Kraidy, Ágota Révész, Pericles Lewis, Marijk van der Wende, Mette Hjort, Jiang Mianheng, Carl Gombrich, Amelia Peterson, Kamal Ahmad, Olga Zlatkin-Troitschanskaia, Fernando Reimers, Stephen M. Kosslyn, Teri A. Cannon, Richard C. Levin, Michael Ignatieff, Emily J. Levine, Katie Abramowitz

The Future of Free Speech
edited by Lee C. Bollinger & Geoffrey R. Stone

The Global Quest for Educational Equity
edited by James A. Banks

Representing the intellectual community in its breadth and diversity, Dædalus *explores the frontiers of knowledge and issues of public importance.*

amacad.org/daedalus